点推定と能動学習を用いた自動単語分割器の分野適応

Graham Neubig　　中田 陽介　　森 信介
京都大学 情報学研究科

1　はじめに

日本語や中国語などの非分かち書き言語では，単語分割は言語処理の第一歩である．最初に行われる処理であるため，単語分割が誤るとその後の処理にも大きく影響する．誤りによる悪影響を防ぐために，単語分割に特に高い精度が求められる．

日本語では，単語分割器（または品詞付与や読み推定も行う形態素解析器）はいくつか公開されている [1, 2]．しかし，その多くは新聞記事などの一般分野での利用を想定しており，専門分野で利用すると精度が低下する．専門分野への適応が必要であるが，多くの場合では本質的に適応は不可能であるか，可能だとしてもコーパスのアノテーションに多くの労力がかかる．

本研究は必要なアノテーション労力の軽減に焦点を置く．重要な箇所のみにタグを付与する部分的アノテーション法を紹介し，このようなタグから学習可能な点推定に基づく単語分割法を採用する．また，アノテーション箇所を選択するための能動学習について述べ，点アノテーションと単語アノテーションの 2 通りのアノテーション戦略を提案する．

それぞれのアノテーション戦略の効果を検証するため，医療分野への適応実験を行った．一般分野のデータのみから始め，従来のフルアノテーションと提案手法によるアノテーションを行い，それぞれの手法の時間効率を測った．その結果，単語アノテーションはフルアノテーションと点アノテーションと同等の時間でより高い分割精度を実現した．

2　単語分割のための言語資源

2.1　単語分割の定義

日本語や中国語などの非分かち書き言語の自然なテキストには明示的な単語境界が含まれていない．つまり，文字列 $X = x_1, x_2, \ldots, x_m$ が分かっても，単語列 $W = w_1, w_2, \ldots, w_n$ が一意に決まるというわけではない．しかし，自然言語処理では原則として単語単位の入力を前提としているため，X を単語に分割して，W を得る必要がある．

近年，このような単語分割は隠れマルコフモデル (HMM) や条件付き確率場 (CRF) など，機械学習に基づいた手法によって行われている [1, 4]．これらの手法を利用する前に，分割器の学習のための言語資源を準備しなければならない．

2.2　入手可能な言語資源

単語分割のために，一般的に以下のような言語資源は入手可能である．

一般分野の分割済みコーパス C_g: 一般分野では京都大学テキストコーパス [8] や現代日本語書き言葉均衡コーパス [3] など，分割済みデータは入手可能となっている．

適応分野の非分割コーパス C_a: 専門分野では分割済みコーパスは必ずしもないが，大規模な非分割データはネット等を通して入手可能な場合が多い．

適応分野の辞書 D_a: 多くの専門分野では人間の参考用に作成された辞書が電子データとして入手可能である．

専門分野のための単語分割器を作成する際，C_g と D_a はそのまま利用できるが，C_a をアノテーションする必要がある．次節では C_a のアノテーション法について述べる．

3　アノテーション法

3.1　フルアノテーション

分割情報を付与するためのもっとも自然な手法は，各単語間に空白を入れることである（ここでは空白を明示的に表示するために「␣」を利用する）．例えば，図 1 の **a** のようなアノテーションなしの文字列が与えられた場合，**b** のように各単語間の間に空白を挿入する．

これは直感的で，アノテーションを行っている作業者に分かりやすいため，ほとんどのアノテーションはこのように行われる．しかし，このようなアノテーション法では，文全体にタグを付与する必要がある．実際のコー

a.	農産物価格安定法を施行
b.	農産␣物␣価格␣安定␣法␣を␣施行
c.	農-産 | 物 | 価-格 | 安-定 | 法 | を | 施-行
d.	農␣産␣物 | 価-格 | 安-定␣法␣を␣施-行

図 1: 各アノテーション法：**a.** なし，**b.** フル，**c.** タグ，**d.** 部分的

パスでは，数百文字の文の中で，アノテーションが必要な箇所が 1 つか 2 つしかない場合が多く，文全体にタグを付与しなくてもフルアノテーションとほぼ同等の精度向上が期待できる．次節は必要な箇所のみにタグを付与するためのアノテーション法を紹介する．

3.2 部分的アノテーション

必要な部分のみをアノテーションするために，まず x_i と x_{i+1} の間に単語境界があるか否かを表すタグ t_i を定義する [7]．t_i は「単語境界あり」を意味する **E** と「単語境界なし」を意味する **N** の 2 値を取り得る．X が与えられた場合，タグ列 $T = t_1, t_2, \ldots, t_{m-1}$ がすべて決まれば，単語列 W も一意に決まる．

実際にアノテーションを行う場合，t_i の値を表す記号を文字 x_i と x_{i+1} の間に挿入する．本稿では，t_i =**E** の場合は「|」を付与し，t_i =**N** の場合は「-」を付与する．この基準でフルアノテーションを行っていけば図 1 の **c.** のようなテキストが得られる．

さらに，t_i の値が未知である場合，「␣」を挿入する．こうすることで，フルアノテーションをする必要がなく，図 1 の **d.** のように必要な箇所のみをアノテーションすることができる．

4 単語分割の推定法

従来の HMM や CRF に基づいた単語分割手法は文ごとに学習されるため，フルアノテーションを前提としており，部分的にアノテーションされたコーパスから学習することは不可能である[1]．部分的アノテーションで得られた情報を活用するために，HMM や CRF のような系列計算ではなく，各タグ t_i の値を個別に推定する点推定を利用する．

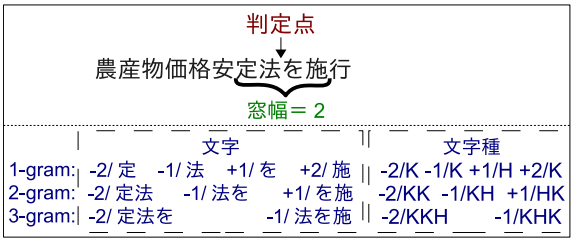

図 2: 窓幅 2 の文字 3-gram と文字種 3-gram 素性

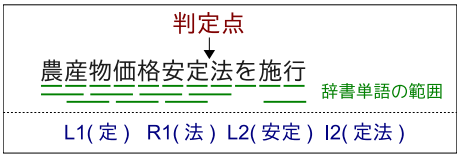

図 3: 辞書単語素性

4.1 点推定の詳細

各タグ t_i は **E** と **N** の 2 通りの値を取り得るため，各タグの推定は 2 値分類問題として扱うことができる [7]．2 値分類問題は機械学習のもっとも一般的な問題の 1 つであり，これらを解く SVM やロジスティック回帰などの洗練した手法は多く存在する．本研究では精度と学習効率の兼ね合いを考慮し，線形 SVM を採用した．

SVM の素性として，以下のものを利用した：

1. **文字 n-gram**：識別するタグ t_i の周りの文字 n-gram を素性として利用する．窓幅 w を決め，$x_{i-w+1}, \ldots, x_i, x_{i+1}, \ldots, x_{i+w}$ の間の文字のみを考慮することで素性の数を抑える（図 2 参照）．

2. **文字種 n-gram**：文字 n-gram と同じ n-gram で，文字自体ではなく，文字の種類（漢字，カタカナ，ひらがな，ローマ字，数字，その他）を利用する．

3. **辞書単語素性**：単語辞書を導入するために，辞書に含まれている単語に関する素性も利用する．これらの素性は，t_i が代表する文字間が長さ k の単語の開始点になってること (R)，長さ k の単語の終了点になっていること (L)，長さ k の単語に含まれていること (I) を表す（図 3 参照）．

これらの素性を抽出し，線形 SVM やロジスティック回帰で単語分割を行う解析器「KyTea」を開発し，オープンソースでリリースした[2]．後述する実験は KyTea を用いて行った．文字 n-gram と文字種 n-gram の n-gram 長と窓幅をすべて 3 とした．

[1] その異例として，Tsuboi らの研究がある [4]．非アノテーション箇所のタグを隠れ変数とし，周辺化を行うことで部分的にアノテーションされたコーパスから CRF を学習する．この手法は効果的ではあるが，アノテーションがスパースであればあるほど周辺化に時間がかかるため，長い文の中で数タグしか存在しない場合は非現実的である．

[2] `http://www.phontron.com/kytea` にて入手可．読み推定の機能も含まれている．

F	前： 農産␣物価␣格安␣定␣法 後： 農産␣物␣価格␣安定␣法
P	前： 農␣産␣物?価␣格␣安␣定␣法 後： 農␣産␣物 | 価␣格␣安␣定␣法
W	前： 農␣産 | 物?価 | 格␣安␣定␣法 後： 農␣産 | 物 | 価-格 | 安␣定␣法

図 4: 各アノテーション戦略 : **F** フルアノテーション，**P** 点アノテーション，**W** 単語アノテーション

5 アノテーション戦略

部分的アノテーションを最大限に活かすためには，単語分割器の精度向上につながる箇所を選択し，それらを部分的にアノテーションする必要がある．本研究では能動学習を用いて交互にアノテーションと分割器の学習を行う．

単語分割器は各タグにつき，信頼度を出力することができる．線形 SVM を利用しているため，各点の SVM 平面からの距離を信頼度として採択する[3]．分割点の信頼度に基づいて，以下のアノテーション戦略を提案する :

フルアノテーション F: C_g で学習された分割器を利用して C_a を分割する．C_a の先頭から分割結果を人手で修正する．

点アノテーション P: C_g で学習された分割器を利用して C_a を分割する．分割結果の中からもっとも信頼度の低い 100 点を選択し，その点 にタグを付与する．C_g と部分的にアノテーションされた C_a から学習された分割器でもう一度 C_a を分割し，タグ付与から繰り返す.

単語アノテーション W: **P** と同じ過程で行われ，その点 だけではなく，その点を開始点か終了点とする単語 または その点を含む単語 にもタグを付与する．

各アノテーション戦略の例を図 4 に示してある．

各戦略の特徴として以下のようなものが考えられる :

F: より少ない時間で多くのタグを付与することが可能であるが，分割器の学習に役立たない箇所も多く含まれている．

P: 各点は確実に分割器の精度向上につながるが，人間でもすぐに分からない点が多い．このため，熟考したり，インターネット等で調べたりする必要があり，1 点につき多くの時間がかかる．

W: **P** と同じく価値ある点をアノテーションし，熟考やインターネットの調査などが必要となる．しかし，**P** と異なり，その情報をすべてアノテーションに入れるため，より多くの箇所がアノテーションされる．

[3] ロジスティック回帰の場合は確率を用いることも可能である．

表 1: コーパスと辞書のデータサイズ

データ	文字数	単語数
一般分野コーパス	1.29M	899k
適応分野コーパス	20.1M	-
テストコーパス	67.8k	45.3k
一般分野辞書	-	223k
適応分野辞書	-	95.3k

6 評価

6.1 実験条件

各アノテーション戦略の効果を検証するために，単語分割器の医療分野への適応実験を行った．

実験に用いたコーパスは，「現代日本語書き言葉均衡コーパス」モニター公開データ（2009 年度版）中の人手による単語分割の修正がなされている文 (一般コーパス) と医療文書からなる適応コーパスである．医療文書のコーパスから 1000 文を取り除き，テストデータとした．さらに，辞書として，UniDic (Ver. 1.3.9)[5] とライフサイエンス辞書 [6] を利用した．コーパスと辞書のサイズは表 1 に示してある．

1 人の作業者が C_a に対して，**F**，**P**，**W** の 3 通りのアノテーションを行った．**P** と **W** の能動学習手法について，100 点ずつアノテーションをしてから，単語分割器の再学習を行い，精度を測った．**F** については，**P** の 100 点と同じ時間作業をしてから，単語分割器の学習を行い精度を測った．作業者の経験量は作業時間に影響を及ぼす恐れがあるため **P**,**F**,**W**,**P**,**F**,**W**... のように，各戦略によるアノテーションを順に繰り替えした．

アノテーションに必要な時間はコストと比例するため，時間効率を評価基準とした．つまり，同等の作業時間で得られたデータから学習されたモデルを対象とし，それぞれの単語分割精度を比較した．単語分割精度の評価基準はタグ正解率とする．

6.2 アノテーション時間

まず，各戦略に必要なアノテーション時間を計った（表 2 参照）．最初の数ターンには作業者がまだアノテーションに慣れておらず，多くの時間がかかるため，最後の 10 ターンのみを対象とする．

タグ付与時間の立場から見ると，**F** は圧倒的に早い．100 点の付与時間では，単純作業がより少ない **P** は少し

表 2: 各戦略のアノテーション時間

戦略	100 点	100 タグ
フルアノテーション **F**	-	16s
点アノテーション **P**	9m15s	9m15s
単語アノテーション **W**	10m35s	2m15s

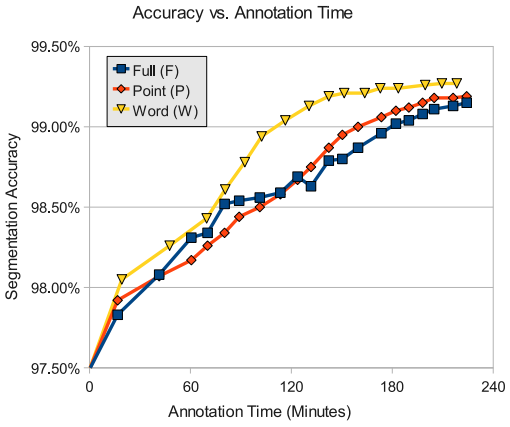

図 5: アノテーション時間と分割精度の関係

早いが，**W** では 1 点に付き平均 4.68 タグを付与したため，タグ付与時間で **W** は **P** を有意に下回る．

6.3 アノテーションの効果

次に，各アノテーション戦略のアノテーション時間と精度向上を測った（図 5 参照）．その結果，**F** と **P** は時間に対してほぼ同等の精度となった．アノテーションタグ数と精度向上の総体関係を表す図 6 から，**P** はもっとも効率の良いタグを選ぶことは明らかであるが，1 点のアノテーションに多くの時間がかかるため，アノテーション量の多い **F** を上回ることはなかった．

その一方，**W** はすべてのアノテーション時間でもっとも良い精度となり，2 時間のアノテーションでは他の戦略との差が特に顕著である．これは能動学習を用いたアノテーション箇所を利用している上で，インターネットで調べた単語境界の情報を無駄にしない効率的なアノテーションが精度に大きく貢献したからである．

7 おわりに

本研究は単語分割の効率的な分野適応のための能動学習と点推定を用いたアノテーション戦略を提案した．点アノテーションは従来のフルアノテーションとほぼ同等

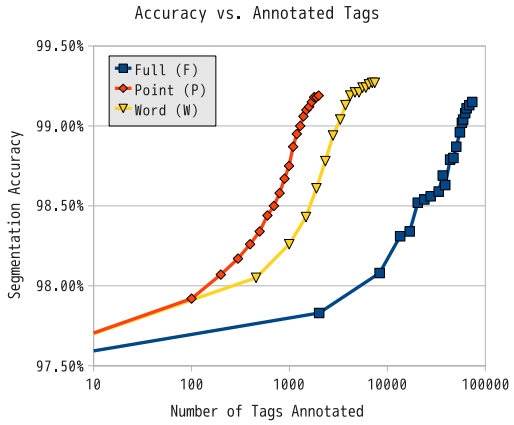

図 6: アノテーションタグ数と分割精度の関係

の精度となり，単語アノテーションは従来の手法を有意に上回った．

これからの課題として，読み推定や品詞推定，固有表現抽出などへの応用がある．また，現在では作業は基本的にテキストエディタで行われているため，インタフェースの改善でさらに効率的なタグ付与が期待できる．

参考文献

[1] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net`, 2009.

[2] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pp. 22–28, 1994.

[3] Kikuo Maekawa. Balanced corpus of contemporary written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102, 2008.

[4] Yuta Tsuboi, Hisashi Kashima, Hiroki Oda, Shinsuke Mori, and Yuji Matsumoto. Training conditional random fields using incomplete annotations. In *Proc. COLING08*, pp. 897–904, 2008.

[5] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–122, 2007.

[6] 金子周司. ライフサイエンス辞書とは. 情報管理, Vol. 49, No. 1, pp. 24–35, 2006.

[7] 森信介, 小田裕樹. 3 種類の辞書による自動単語分割の精度向上. 情処研報 NL-193, 2009.

[8] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言処第 3 回年次大会, pp. 115–118, 1997.