3種類の辞書による自動単語分割の精度向上

森 信介[†]・小田 裕樹

概要

本論文では、日本語の文の自動単語分割をある分野に適用する現実的な状況において、精度向上を図るための新しい方法を提案する。提案手法の最大の特徴は、複合語を参照することが可能な点である。複合語とは、内部の単語境界情報がなく、その両端も自動分割器の学習コーパスの作成に用いられた単語分割基準と必ずしも合致しない文字列である。このような複合語は、自然言語処理をある分野に適用する多くの場合に、利用可能な数少ない言語資源である。提案する自動単語分割器は、複合語に加えて単語や単語列を参照することも可能である。これにより、少ない人的コストでさらなる精度向上を図ることが可能である。

実験では、これらの辞書を参照する自動単語分割システムを最大エントロピー法を用いて構築し、それぞれの辞書を参照する場合の自動単語分割の精度を比較した。実験の結果、本論文で提案する自動単語分割器は、複合語や単語列を参照することにより、対象分野においてより高い分割精度を実現することが確認された。

Automatic Word Segmentation using Three Types of Dictionaries

Shinsuke MORI[†] and Hiroki Oda

Summary

In this paper we propose a new method for automatically segmenting a sentence in Japanese into a word sequence. The main advantage of our method is that the segmenter is, by using a maximum entropy framework, capable of referring to a list of compound words, i.e. word sequences without boundary information. This allows for a higher segmentation accuracy in many real situations where only some electronic dictionaries, whose entries are not consistent with the word segmentation standard, are available. Our method is also capable of exploiting a list of word sequences. It allows us to obtain a far greater accuracy gain with low manual annotation cost.

We prepared segmented corpora, a compound word list, and a word sequence list. Then we conducted experiments to compare automatic word segmenters referring to various types of dictionaries. The results showed that the word segmenter we proposed is capable of exploiting a list of compound words and word sequences to yield a higher accuracy under realistic situations.

[†]京都大学 学術情報メディアセンター, Kyoto University, Academic Center for Computing and Media Studies

1　はじめに

日本語や中国語のように、明示的な単語境界がない言語においては、自動単語分割は自然言語処理の最初のタスクである。ほとんどの自然言語処理システムは、単語単位に依存しており、自動単語分割器はこれらの言語に対して非常に重要である。このような背景の下、人手による単語分割がなされた文からなるコーパスを構築する努力 (黒橋，長尾 1997; Maekawa 2008) とともに、経験的手法による自動単語分割器や同時に品詞を付与する形態素解析器の構築 (永田 1999; 森，長尾 1998; 小田，森，北 1999; Sproat and Chang 1996; 内元，関根，井佐原 2001; 工藤，山本，松本 2004) が試みられてきた。

近年、自然言語処理が様々な分野に適用されている。特許開示書の自動翻訳、裁判記録の自動作成のための音声認識用の言語モデル作成、医療文章からの情報抽出などである。これらの応用では品詞は不要なので、本論文では品詞を付与しない単語分割を扱う。単語分割では、コーパス作成の労力を単語境界付与に集中することができるので、品詞付与が必要となる形態素解析を前提とするよりもより実用的であることが多い。現在の自動単語分割器は、一般的な分野のコーパスから構築されており、上述のような様々な分野の文を高い精度で単語分割できない。とりわけ、対象分野特有の単語や表現の周辺での精度の低下が著しい。これらの対象分野に特有の単語や表現は、処理すべき文において重要な情報を保持しているので、この問題は深刻である。

このような問題を解決するためには、対象分野での単語分割精度の向上を図る必要がある。理想的方法は、ある程度の量の対象分野の文を、一般分野のコーパス作成と同じ単語分割基準に沿って人手で単語分割し、自動単語分割器を再学習することである。しかしながら、多くの実際の状況では、人による利用を想定した辞書が対象分野の唯一の追加的言語資源である。これらの見出し語は、単語分割基準とは無関係に選定されており、単語分割基準に照らすと単語ではないことが多い。本論文では、これらの見出し語のように、内部の単語分割情報が与えられておらず、かつ両端が単語境界であるという保証がない文字列を複合語と呼ぶ。本論文では、単語分割済みコーパスに加えて、複合語辞書を参照する自動分割器を提案する。ほとんどの複合語は両端が単語境界であり、内部に単語分割基準に従って単語境界情報を付与することで単語列に変換することが可能である。このために必要な人的コストは、適用分野の単語分割済みコーパスの作成に比べて非常に少ない。本論文ではさらに、単語列辞書を参照し精度向上を図る自動単語分割器を提案する。提案手法を用いることにより、一般に販売されている辞書 (複合語辞書) を参照することで、付加的な人的コストなしに、ある分野における自動単語分割の精度を向上させることができる。また、単語列辞書を参照する機能により、コーパスを準備するよりもはるかに低い人的コストでさらなる精度向上を実現することが可能になる。

2 単語分割のための言語資源

この節では、まず、日本語を例に単語境界を明示しない言語の文を単語に分割する問題について説明する。次に、入力文を自動的に単語列に分割する自動単語分割器を構築するために利用可能な言語資源について述べる。

2.1 単語分割問題

日本語や中国語のように、単語境界を明示しない言語は多数ある。これらの言語における自然言語処理の第一歩は、文に単語境界情報を付与することである。以下では、次の文を入力の例として、単語分割問題を説明する。

入力: 畜産物価格安定法を施行

単語分割問題は、入力の文字列の全ての隣接する 2 文字間に単語境界を置くか置かないかを決定する 2 値分類問題である。注目している 2 文字が異なる単語に属する場合には空白を、同じ単語に属する場合には何も置かないとすると、上記の例文を適切に単語に分割した結果は以下のようになる。

出力: 畜産 物 価格 安定 法 を 施行

このように、入力文字列を単語に分割することで、単語を単位とした自然言語処理技術を適用することが可能となる。ただし、単語分割における誤りは、後続する自然言語処理の精度を低下させる点に注意しなければならない。

2.2 単語分割済みコーパス

自動単語分割器に関する研究は多数あり、そのほとんどがデータに基づく方法を採用している (永田 1999; 森，長尾 1998; Sproat and Chang 1996; 内元他 2001; 工藤他 2004)[1]。これらの自動単語分割器は、単語分割基準に従って人手で単語に分割されたコーパスからパラメータを推定する。したがって、学習コーパスにおける誤りは自動分割の結果に波及し、後続する自然言語処理のアプリケーションの精度を損なう。それゆえ、単語分割済みコーパスの質は、非常に重要である。高い分割精度を確保するためにも、後続する自然言語処理を適用する対象分野の単語分割済み文を自動単語分割システムの学習コーパスとすることが望ましい。しかしながら、単語分割基準に従って正確に単語分割されたコーパスを用意するコストは非常に高い。というのも、作業者は、対象分野の用語と単語分割基準を熟知している必要があるからである。実際、コンピューター操作に熟練し単語分割基準を熟知した作業者が、ある分野の 5,000 文を非常に高い精度で単語分割するのに 2 週間 (8 時間 ×10 日) を要したという事実もある[2]。したがって、低いコストで準備できる言語資源のみを

[1] 単語分割と同時に品詞を付与する形態素解析器は、品詞を無視することで自動単語分割器として用いることができるので、自動単語分割器の研究に含めている。

[2] これは著者の実際の経験に基づいている。その際の作業においては、高機能のエディターを高度に利用し、必要に応じてプログラムを記述・実行し、自動分割器を適宜再構築することで単語分割誤りを非常に効率良く修正した。

| ： 単語境界がある。

- ： 単語境界がない。

␣ ： 単語境界の有無は不明である。

図 1　単語境界の 3 値表現

用いて対象分野の文を高い精度で分割する自動単語分割器の実現方法は非常に有用である。

2.3　3 種類の辞書

単語分割基準が所与とすれば、単語分割問題に用いることができる辞書は、3 つに分類できる。以下では、これら 3 種類の辞書について、図 1 に示した 3 値表現を用いて詳述する。

- **単語辞書:**

  この辞書は、単語分割基準に従う単語からなる。つまり、この辞書に含まれる文字列は、ある文脈で最左の文字の左と最右の文字の右に単語境界があり、さらにその内部には単語境界がない。例えば、3 値表現された以下の文字列は単語である。

  |言-語|

- **単語列辞書:**

  この辞書は、単語の列からなる。つまり、この辞書に含まれる文字列は、ある文脈で最左の文字の左と最右の文字の右に単語境界があり、さらに、各文字列のすべての文字間に単語分割基準に従って単語境界情報が付与されている。例えば、3 値表現された以下の文字列は単語列である。

  |計-算|言-語|学|

- **複合語辞書:**

  この辞書は、単語の連接である文字列からなる。つまり、この辞書に含まれる文字列は、ある文脈でその左右両端に単語境界があるが、文字列内部の単語境界情報は不明である。例えば、3 値表現された以下の文字列は複合語である。

  |計␣算␣言␣語␣学|

人が利用することを想定した商用・非商用の機械可読辞書は多数ある。実際、様々な対象分野における専門用語や固有名詞を多数含む辞書がある (日外アソシエーツ 2001)。また、仮名漢字変換のための辞書が様々な分野において公開されている (金子 2003)。これらの辞書の見出し語は、自動単語分割器が学習に用いるコーパスの単語分割基準とは無関係に選定されており、多くの自動単語分割器において、これを精度向上に直接利用することはできない。

単語分割基準に照らすと、人が利用することを想定した辞書の見出し語の多くは、上記の分類では複合語である。左右両端は単語境界であることがかなりの確度で期待できるが、文字列の内部の単語境界情報がない。複合語は、人的コストをかけて単語列に変換することができる。この際に必

要な作業は、左右の両端が単語境界であることのチェックと、文字列内のすべての文字境界に単語境界情報を付与することである。このコストは、対象分野の単語分割済みコーパスの作成に要するコストに比べて非常に少ない。以上の議論から、複合語や単語列を参照することで精度が向上する自動単語分割器を構築することは実用的意義が非常に大きい。

3 単語分割法

本節では、前節の 3 種類の辞書を学習データとする日本語単語分割法について述べる。提案手法は、3 種類の辞書と単語分割済みコーパス (部分的にアノテーションされていれば良い) を学習データとする。

3.1 最大エントロピーモデルによる点予測単語分割

日本語の単語分割の問題は、入力文の各文字間に単語境界が発生するか否かを予測する問題とみなせる (小田他 1999; 颯々野 2006; 風間，宮尾，辻井 2004; Tsuboi, Kashima, Mori, Oda, and Matsumoto 2008)。つまり、文 $\boldsymbol{x} = x_1 x_2 \cdots x_m$ に対して、文字 x_i と x_{i+1} の間が単語境界であるか否かを表すタグ t_i を付与する問題とみなせる。

付与するタグは、単語境界であることを表すタグ $\mathbf{E}$ (“E” に相当) と、非単語境界であることを表すタグ $\mathbf{N}$ (“N” に相当) の 2 つのタグからなる。各文字間のタグを単語境界が明示されたコーパスから学習された最大エントロピーモデル (ME model; maximum entropy model) により推定する[3]。その結果、より高い確率を与えられたタグをその文字間のタグとし、単語境界を決定する。すなわち、以下の式が示すように、最大エントロピーモデルにより、単語境界と推定される確率が非単語境界と推定される確率より高い文字間を単語境界とする。

$$P_{ME}(\mathbf{E}|i, \boldsymbol{x}) > P_{ME}(\mathbf{N}|i, \boldsymbol{x})$$

これにより、入力文を単語に分割することができる。

最大エントロピーモデルによる単語分割法では、単語境界情報が付与された $\boldsymbol{x} = x_1 x_2 \cdots x_m$ の各文字間を、タグ t_i と素性の組合せ S とみなして、学習と確率推定を行う。

$$S = \{(t_i,\, f_{i,1}(\boldsymbol{x}),\, f_{i,2}(\boldsymbol{x}),\, \cdots) \mid 1 \leq \forall i \leq m-1\},$$

素性 $f_{i,j}(\boldsymbol{x})$ の詳細は次項で述べる。

[3]文献 (坪井，森，鹿島，小田，松本 2009) のように CRF(conditional random fields) により推定することもできるが、計算コストと記憶領域が大きくなる。これらの差は、スパースな部分的アノテーションコーパスからの学習において顕著となる。つまり、CRF のように系列としてモデル化する方法では、アノテーションのない部分も考慮する必要があるのに対して、点推定の最大エントロピーモデルでは、アノテーションのある部分のみを考慮すればよい。このような考察から、本論文では計算コストの少ない最大エントロピーモデルを用いる。

3.2　参照する素性

文字 x_i と x_{i+1} の間に注目する際の最大エントロピーモデルの素性としては、文字列 x_{i-1}^{i+2} に含まれるの部分文字列である文字 n-gram および字種 n-gram ($n = 1, 2, 3$) をすべて用いる[4]。ただし、以下の点を考慮している。

- 素性として利用する n-gram は、先頭文字の字種がその前の文字の字種と同じか否か、および、末尾文字の字種がその次の文字の字種と同じか否かの情報を付加して参照する[5]。
- 素性には注目する文字間の位置情報を付加する。

3.3　辞書の利用

さらに、前述した 3 種類の辞書を参照し、以下の素性を用いることを提案する。

- 文字列 x_{i-1}^{i+2} に含まれる文字 n-gram ($n = 1, 2, 3$) が、単語辞書中の単語と一致する文字列であるか否かを表す 9 素性 (9 つの位置の文字 n-gram について判定)
- 注目する文字境界 ($x_i x_{i+1}$ の間) の辞書中の位置を表す以下の 4 素性
  - 単語の開始位置の “|” に該当するか否かを表す素性 (単語、単語列、複合語のいずれかが $x_{i+1}x_{i+2}\cdots$ に前方一致するか否か)
  - 単語の終了位置の “|” に該当するか否かを表す素性 (単語、単語列、複合語のいずれかが $\cdots x_{i-1}x_i$ に後方一致するか否か)
  - 単語列の単語境界の “|” に該当するか否かを表す素性 ($\cdots x_i x_{i+1} \cdots$ の位置にいずれかの単語列が出現し、かつ x_i と x_{i+1} の間が単語列中の “|” に該当するか否か)
  - 単語や単語列の “–” に該当するか否かを表す素性 ($\cdots x_i x_{i+1} \cdots$ の位置にいずれかの単語か単語列が出現し、かつ x_i と x_{i+1} の間が “–” に該当するか否か)

4　評価

提案手法の評価のために、様々な自動単語分割器を構築し、テストコーパスに対する分割精度を測定した。この節では、その結果を提示し提案手法の評価を行う。

4.1　実験条件

まず、対象分野のテストコーパスを日本経済新聞 (日本経済新聞社 2001) の記事とした。学習コーパスは一般分野の単語分割済みコーパスである。評価のために、学習コーパスと同じ分野のテストコーパスも用いた (表 1 参照)。一般分野コーパスは、現代日本語書き言葉均衡コーパス (Maekawa

[4]字種は、漢字、ひらがな、カタカナ、アルファベット、数字、記号の 6 つとした。

[5]パラメータ数の急激な増加を抑えつつ素性の情報量を増加させることを意図している。これにより、参照範囲を前後 1 文字拡張して x_{i-2}^{i+3} の範囲の n-gram ($n = 3, 4, 5$) が参照されることになる。

表 1　単語分割済みコーパス

分野	用途	文数	単語数	文字数
一般	学習	27,935	626,700	878,089
一般	テスト	3,447	77,990	109,064
新聞記事	テスト	9,023	263,427	398,570

2008)(13,181 文) と、日常会話のための辞書の例文 (ドナルド，羽鳥，山田，伊良部 1992)(14,754 文) である。すべての文は、人手により適切に単語に分割されている。実験では、9-fold の交差検定を行った。つまり、テストコーパスを 9 つの部分に分割し、8 つの部分を複合語や単語列の選定に用い、残りの 1 つを自動分割のテストとして用いることを、9 通りに渡って行った。実験に際して、2 節で述べた 3 種類の辞書を用意した。1 つ目は単語辞書 (UniDic-1.3.9) で、語彙サイズが非常に大きいことに加えて、その見出し語が学習コーパスと同じ単語分割基準に従っていることが注意深くチェックされている (伝，小木曽，小椋，山田，峯松，内元，小磯 2007)。2 つ目は複合語辞書で、その見出し語は機械可読の商用辞書 (日外アソシエーツ 2001) から得た。その多くは、専門用語と固有名詞である。実験では、テストとして用いられる 1 つの部分コーパス以外の 8 つの部分コーパスに文字列として出現する複合語を用いた。これらの複合語は、両端は単語分割基準と一致していることが期待されるが、その保証はない。3 つ目は単語列辞書である。単語列辞書は、複合語辞書の見出し語を単語分割基準に従って人手により分割することで得られる。単語列は、結果的に 1 単語である場合もある。実験では、8 つの部分コーパスに文字列として出現する単語列を用いた。複合語の左右どちらかの端が単語分割基準と一致していない場合は単語列辞書から除外した。したがって、単語列の数は複合語の数よりも少なくなっている。

表 2 にこれらの辞書の諸元を示す。この表から複合語と単語列の平均文字数はそれぞれ 3.04 文字と 3.12 文字であることが分かる。これらは、学習コーパスの単語長 (1.40 文字) や新聞コーパスの平均単語長 (1.50 文字) より長い。単語列は平均 1.61 単語からなる。自動単語分割器のパラメータは、学習コーパスとこれらの辞書を同時に参照し推定される。

4.2　評価基準

自動単語分割の評価に用いた基準は、適合率、再現率、境界推定精度及び文正解率である。これらの計算方法を以下のような例を用いて説明する。ここで、自動単語分割の結果を AWS、正解の単語列を COR としている。

AWS: 畜産 物価 格安 定法 を 施行
COR: 畜産 物 価格 安定 法 を 施行

表 2　辞書

ID	種類	分野	見出し数	単語数	文字数
$\boldsymbol{w1}$	単語	一般分野	145,310.0	145,310.0	430,797.0
$\boldsymbol{w2}$	単語	一般分野	27,466.0	27,466.0	81,433.0
$\boldsymbol{s}$	単語列	適応分野	17,099.5	27,465.3	53,364.4
$\boldsymbol{c}$	複合語	適応分野	19,697.1	–	59,868.4

※ $\boldsymbol{w2}$ は $\boldsymbol{w1}$ から 27,466 語を無作為抽出した結果であり、単語列と複合語の値は交差検定の平均である。

境界推定精度は、単語境界情報が正しく推定された文字間の割合である。上記の例では、文字数は 11 あるので、単語境界の推定対象となる文字間の数は 10 である。この内、単語境界情報が正しく推定された文字間は 5 であるので境界推定精度は 5/10 となる。文正解率は、すべての文字間において単語境界情報が正しく推定された文の割合である。適合率と再現率は、以下のように計算される。まず、正解の単語列に含まれる単語数を N_{COR}、自動単語分割の結果に含まれる単語数を N_{AWS} とし、さらに正解の単語列と自動単語分割の結果の最長部分一致単語列に含まれる単語数を N_{LCS} とする。この定義の下、適合率は N_{LCS}/N_{AWS} で与えられ、再現率は N_{LCS}/N_{COR} で与えられる。上述の例では、最長部分一致単語列は下線を引いた単語列であり、その単語数から $N_{LCS} = 3$ である。正解の単語列の単語数から $N_{COR} = 7$ であり、自動単語分割の結果の単語数から $N_{AWS} = 6$ である。したがって、再現率は $N_{LCS}/N_{COR} = 3/7$ であり、適合率は $N_{LCS}/N_{AWS} = 3/6$ である。

4.3　評価

参照する辞書による単語分割精度の差を調べるために、辞書を参照せずコーパスのみから学習する自動単語分割器 $\boldsymbol{B}$(ベースライン) を構築し、さらに以下の 4 つの自動単語分割器を構築した。

$\boldsymbol{W1}$: コーパスに加えて単語辞書 $\boldsymbol{w1}$ を参照

$\boldsymbol{W2}$: コーパスに加えて単語辞書 $\boldsymbol{w2}$ を参照

$\boldsymbol{S}$: コーパスに加えて単語列辞書 $\boldsymbol{s}$ を参照

$\boldsymbol{C}$: コーパスに加えて複合語辞書 $\boldsymbol{c}$ を参照

これらの自動単語分割器による一般分野における分割精度を表 3 に、対象分野における分割精度を表 4 に示す。

表 3 から、ベースラインである自動単語分割器 $\boldsymbol{B}$ の一般分野における分割精度は十分高いが、表 4 から、対象分野においては分割精度が著しく低下することがわかる。自動単語分割器 $\boldsymbol{C}$ の結果から、複合語辞書を用いることで対象分野における分割精度が向上することが分かる。複合語辞書としては、多くの分野において利用可能な人のための辞書を直接用いることができるので、付加的な

表 3　一般分野における自動単語分割の精度

ID	学習に用いた資源	境界推定精度	適合率	再現率	文精度
$\boldsymbol{B}$	単語分割済みコーパス	98.82%	97.87%	97.86%	77.22%
$\boldsymbol{C}$	＋ 複合語辞書 $\boldsymbol{c}$	98.86%	97.93%	97.91%	77.62%
$\boldsymbol{S}$	＋ 単語列辞書 $\boldsymbol{s}$	98.97%	98.08%	98.16%	78.97%
$\boldsymbol{W1}$	＋ 単語辞書 $\boldsymbol{w1}$	98.99%	98.13%	98.13%	79.12%
$\boldsymbol{W2}$	＋ 単語辞書 $\boldsymbol{w2}$	98.84%	97.88%	97.89%	77.22%

表 4　適用分野における自動単語分割の精度

ID	学習に用いた資源	境界推定精度	適合率	再現率	文精度
$\boldsymbol{B}$	単語分割済みコーパス	98.07%	96.28%	96.28%	55.64%
$\boldsymbol{C}$	＋ 複合語辞書 $\boldsymbol{c}$	98.20%	96.44%	96.50%	56.95%
$\boldsymbol{S}$	＋ 単語列辞書 $\boldsymbol{s}$	98.72%	97.39%	97.47%	65.54%
$\boldsymbol{W1}$	＋ 単語辞書 $\boldsymbol{w1}$	98.43%	96.93%	96.88%	60.81%
$\boldsymbol{W2}$	＋ 単語辞書 $\boldsymbol{w2}$	98.09%	96.30%	96.31%	55.95%

人的コストを必要としない。このことから、複合語辞書を参照することで自動単語分割器の精度向上が実現できることは非常に有用であるといえる。

自動単語分割器 $\boldsymbol{C}$ の分割精度は、両分野において、コーパスと同じ基準で単語に分割された単語辞書 $\boldsymbol{w1}$ を参照する自動単語分割器 $\boldsymbol{W1}$ の分割精度より低い。これは、単語辞書 $\boldsymbol{w1}$ の見出し語の数が、約 14.5 万語と非常に大きいことと、本実験での適応対象である新聞記事が、単語辞書の想定分野となっていることによる。このように、大きな単語辞書を、様々な分野に対して準備することは現実的ではないであろう。実際、単語辞書に含まれる単語の合計の文字数は 430,797 文字 (表 2 参照) と非常に大きく、これは、対象分野の約 9,879 文 (表 1 参照) に相当する。2.2 項で述べたように、非常に熟練した作業者の単語分割速度が 1 日 500 文程度であるから、この量のコーパスを作成するには、対象分野の表現と単語分割基準を熟知した作業者が約 20 日間作業にあたる必要があると考えられる。

単語列辞書 $\boldsymbol{s}$ と同じ単語数の単語辞書 $\boldsymbol{w2}$ を用いる自動単語分割器 $\boldsymbol{W2}$ の分割精度を自動単語分割器 $\boldsymbol{C}$ と比べると、$\boldsymbol{C}$ の精度は $\boldsymbol{W2}$ よりも高い。複合語辞書は、処理すべき適用分野のテキストと共に提供されることが多い。これらのことから、ある分野に自動単語分割器を適用する場合、

一般的な分野の辞書を整備するのではなく、その分野の複合語辞書を利用するのがよい戦略であるといえる。

単語列辞書を参照する自動単語分割器 $\boldsymbol{S}$ の対象分野における単語分割精度は、複合語辞書を参照する自動単語分割器 $\boldsymbol{C}$ よりも高い。これは、単語列辞書が複合語辞書に単語境界情報を付与した結果であることを考えると当然である。このことから、対象分野のテキストに出現する複合語に単語境界情報を付与することで、さらなる精度向上を実現できることが分かる。

文字列「共同宣言」を複合語としてまたは単語列として辞書に追加することにより分割精度が向上した例を次に示す。

$\boldsymbol{B}$:　…|日-米|安-保|共|同|宣-言-取|り|ま-と-め|…

$\boldsymbol{C}$:　…|日-米|安-保|共|同|宣-言|取|り|ま-と-め|…

$\boldsymbol{S}$:　…|日-米|安-保|共-同|宣-言|取|り|ま-と-め|…

自動単語分割器 $\boldsymbol{B}$ では「宣言取」を単語であると出力しているが、$\boldsymbol{C}$ では、複合語「|共␣同␣宣␣言|」を参照することで文字列「共同宣言」の後に単語境界があることが分かり、正しく「宣言」と「取」に分割されている。さらに $\boldsymbol{S}$ では、単語列「|共-同|宣-言|」を参照することですべての箇所で正しく単語に分割されている。なお、分野特有であると思われる表現でも、実験に利用した商用辞書に含まれていない単語列に関しては、辞書の追加による精度向上はみられなかった。

自動単語分割器 $\boldsymbol{S}$ による対象分野の分割精度は、大規模な単語辞書を参照する自動単語分割器 $\boldsymbol{W1}$ による分割精度よりも高い。複合語辞書に含まれる合計の文字数は、59,868 文字 (表 2 参照) で、これは、対象分野の 1,373 文に相当する。アノテーションに必要な人的コストは、単語辞書の構築に必要な人的コストと比べて非常に少ない。さらに、複合語の多くが専門用語と固有名詞であるので、作業者に要求される技能は、分野知識と名詞に関する単語分割基準の熟知である。したがって、作業者の確保はコーパスの準備に比べてはるかに容易である。

以上のような考察から、現実的な状況において、既存の辞書のカバー率が高くないある対象分野の自動単語分割器を構築する最良の戦略は、

(1)　対象分野のテキストに出現する複合語辞書の見出し語を収集し、提案する自動単語分割器を用いる

(2)　作業者が確保できる場合には、さらにこれらの複合語に単語境界情報を付与し、提案する自動単語分割器を用いる

であると結論できる。

5　関連研究

自動単語分割の問題をある文字の次に単語境界があるかの予測として定式化することはかなり以前から行われている (永井，日高 1993; 山本，増山 1997; 小田他 1999)。これらの研究では、文字

に単語境界情報を付与して予測単位としている。文字間に単語境界があるかを識別学習により決定することも提案されている (颯々野 2006)。この研究の主眼は能動学習の調査・分析である。辞書の利用に関する記述はなく、また分野適応についても述べられていない。

統計的手法による日本語の文の自動単語分割に関する初期の研究は、丸山ら (丸山，荻野，渡辺 1991) による単語 n-gram モデルを用いる方法がある。また、永田 (永田 1999) による品詞 n-gram モデルによる形態素解析[6]もある。森ら (森，長尾 1998) は、すべての品詞を語彙化した形態素 n-gram モデルを用いることによる精度向上を報告し、さらに単語辞書の参照を可能にする方法を提案し、それによる精度向上を報告している。内元ら (内元他 2001) は、最大エントロピーモデルを用いる形態素解析において単語辞書を参照する方法を提案している。このように、単語分割基準に沿った単語辞書を参照する方法はすでにあるが、複合語や単語列を参照し精度向上を実現した自動単語分割器の報告は、我々の知る限りない。

なお、提案手法は、形態素解析にも拡張可能である。提案する自動単語分割器は音声認識や仮名漢字変換の言語モデル作成に用いることを想定しているので、品詞を推定する必要はないと考えている。品詞も推定すべきか、どの程度の粒度の品詞体系にすべきか、などの問題は後続する自然言語処理と準備すべきデータの作成コストを含む全体の問題であり、本論文の議論を超える。

複合語や単語列を参照し精度向上を図る取り組みは、完全に単語に分割された文に加えて、それ以外の断片的な情報を用いて精度向上を図る取り組みの一つである。坪井ら (坪井他 2009) は、学習コーパスの文の単語境界情報が部分的であるような不完全なアノテーションからも条件付き確率場による自動単語分割器や形態素解析器を学習できる枠組みを提案している。本論文で提案する自動単語分割器は点予測を用いているので、単語分割に関してはこの問題は解決されているといえる。したがって、提案する自動単語分割器は、単語境界情報が部分的に付与されたコーパスと複合語や単語列のすべてを同時に参照することができる。

自動単語分割は、中国語においても提案されている (Sproat and Chang 1996)。自動単語分割器は、空白で区切られる単位が大きい韓国語やフィンランド語などにおいても有用で、言語モデルの作成 (Kwon and Park 2003; Hirsimäki, Creutz, Siivola, Kurimo, Virpioja, and Pylkkönen 2006) に用いられている。提案手法は、これらの言語に対する言語処理においても有用である。

6 おわりに

本論文では、日本語の文の自動単語分割器をある分野に適用する現実的な状況において、精度向上を図るための新しい方法を提案した。提案手法の最大の特徴は、複合語を参照することが可能な点である。本論文で言う複合語とは、内部の単語境界情報がない文字列であり、人の利用を想定し

[6] ここでは、自動単語分割に加えてそれぞれの単語の品詞を同時に推定する処理を形態素解析と呼んでいる。

た辞書の見出し語の多くが複合語である。この機能により、一般に販売されている辞書を参照し、付加的な人的コストなしに、ある分野における自動単語分割の精度を向上させることができる。提案する自動単語分割器は、内部に単語境界情報をもつ単語列を参照することも可能である。この機能により、コーパスを準備するよりも低い人的コストでさらなる精度向上を実現することが可能になる。

実験では、これらの辞書を参照する自動単語分割器を最大エントロピー法を用いて構築し、これらのさまざまな辞書を参照する場合の自動単語分割の精度を比較した。実験の結果、本論文で提案する自動単語分割器は、複合語を参照することにより、より高い分割精度を人的コストなしに実現することが確認された。また、単語列を参照することにより、少ない人的コストでさらなる精度向上が実現されることが示された。したがって、本論文で提案する自動単語分割器は、自然言語処理をある分野に適用する場合に非常に有用である。

参考文献

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. (2006). "Unlimited vocabulary speech recognition with morph language models applied to Finnish." *Computer Speech and Language*, **20**, pp. 515–541.

Kwon, O.-W. and Park, J. (2003). "Korean large vocabulary continuous speech recognition with morpheme-based recognition units." *Speech Communication*, **39**, pp. 287–300.

Maekawa, K. (2008). "Balanced Corpus of Contemporary Written Japanese." In *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102.

Sproat, R. and Chang, C. S. W. G. N. (1996). "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese." *Computational Linguistics*, **22** (3), pp. 377–404.

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., and Matsumoto, Y. (2008). "Training Conditional Random Fields Using Incomplete Annotations." In *Proceedings of the 22th International Conference on Computational Linguistics*.

ドナルドキーン，羽鳥博愛，山田晴子，伊良部祥子 (1992). "会話作文英語表現辞典." 朝日出版社.

颯々野学 (2006). "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, **13** (2), pp. 27–41.

山本幹雄，増山正和 (1997). "品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析." 言語処理学会第 3 回年次大会発表論文集, pp. 421–424.

丸山宏，荻野紫穂，渡辺日出雄 (1991). "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.

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 (2007). "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, **22**, pp. 101–122.

金子周司 (2003). “無料ライフサイエンス辞書の活用と効能.” ファルマシア, **42** (5), pp. 463–467.

永井秀利，日高達 (1993). “日本語における単語の造語モデルとその評価.” 情報処理学会論文誌, **34** (9), pp. 1944–1955.

風間淳一，宮尾祐介，辻井潤一 (2004). “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, **11** (4), pp. 3–24.

永田昌明 (1999). “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, **40** (9), pp. 3420–3431.

森信介，長尾眞 (1998). “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, **5** (2), pp. 75–103.

内元清貴，関根聡，井佐原均 (2001). “最大エントロピーモデルに基づく形態素解析–未知語の問題の解決策–.” 自然言語処理, **8** (1), pp. 127–141.

工藤拓，山本薫，松本裕治 (2004). “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.

黒橋禎夫，長尾眞 (1997). “京都大学テキストコーパス・プロジェクト.” 言語処理学会第 3 回年次大会発表論文集, pp. 115–118.

日外アソシエーツ (2001). “CD-科学技術 45 万語対訳辞典 英和／和英.”.

日本経済新聞社 (2001). “日経全文記事データベース [4 紙版].”.

坪井祐太，森信介，鹿島久嗣，小田裕樹，松本裕治 (2009). “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, **50** (6), pp. 1622–1635.

小田裕樹，森信介，北研二 (1999). “文字クラスモデルによる日本語単語分割.” 自然言語処理, **6** (7), pp. 93–108.