D-bigramを用いた形態素解析処理からの複合語抽出

倉田 岳人，岡崎 直観，石塚 満
東京大学大学院 情報理工学系研究科 電子情報学専攻

1 はじめに

近年，様々な情報が電子化され，我々はこれらを利用することができる．しかし，英語等と異なり，日本語は分かち書きを行わない，このため，電子化された情報を処理する場合，形態素解析を行う必要が生じる．現在幅広く用いられている形態素解析システムは，辞書に基づいた処理を基本としている．しかし，辞書に基づいた手法では，形態素解析の精度が辞書の精度に依存することになる．すなわち，辞書に登録されていない未知語や，既存の単語の組合わせによる新しい単語の処理を行うことができない．これは，日々新しい単語が生まれる現在の情報処理としては問題があると考えられる．

本稿では，辞書を用いない文字間のD-bigramを用いた統計量を用いて文字間のつながりの強さを算出し，形態素解析の結果の修正を行う手法を提案する．

2 形態素解析の仕組みと問題点

形態素辞書に基づく形態素解析の概要を以下に示す．

1. テキスト中のある特定の位置から始まるすべての可能な形態素を辞書引きにより得る．

2. 得られた個々の形態素に対して，形態素コスト及び品詞間の連接コストの和が最小となるパスを算出する．

上記の様な辞書に基づく形態素解析を行った場合，未知語が解析対象テキスト中に現れた時点で解析に失敗する可能性が高い．また，複合語の場合は過分割が生じやすくなることも明らかである．実際，未知語や複合語を含む文に対する処理においては，過分割が生じやすいことが報告されている [1]．

Compound Words Extraction with D-bigram Statistics in Japanese Morphological Analysis

Gakuto KURATA, Naoaki OKAZAKI, Mitsuru ISHIZUKA
Dept. of Information and Communication Engineering, Graduate School of Information Science and Technology, University of Tokyo

3 提案する手法

3.1 隣接文字間の繋がりの強さの評価

文字D-bigramを用いて隣接する文字間の繋がりの強さを評価する手法が提案されている [2]．一般的な文字bigramは隣接する文字の関係のみを扱うが，文字D-bigramでは数文字離れた文字間の共起の関係も考慮する．[2] では，文中の i 番目の文字 w_i と $i+1$ 番目の文字 w_{i+1} の繋がりの強さ $S(i)$(以下,「隣接スコア」と呼ぶ．) を以下の様に定義している．

$$S(i) = \sum_{d=1}^{d_{max}} \sum_{j=i-(d-1)}^{i} MI_d(w_j, w_{j+d}; d) \times g(d)$$

ここで MI_d として，2文字間の相互情報量を文字D-bigramに対応するように拡張したものであり，式 (1) で表される．

$$MI_d(x, y; d) = \log \frac{P(x, y; d)}{P(x)P(y)} \tag{1}$$

また $g(d) = 1/d^2$，$d_{max} = 5$ とした．

3.2 ドメイン固有の文字列の情報

あるドメインのコーパスから，式 (1) の $P(x, y; d)$，$P(x)$，$P(y)$ を算出することができる．その結果を用いて，解析対象の文に対して，図1の様な評価を得ることができる．ここで，縦軸は隣接スコアを表している．

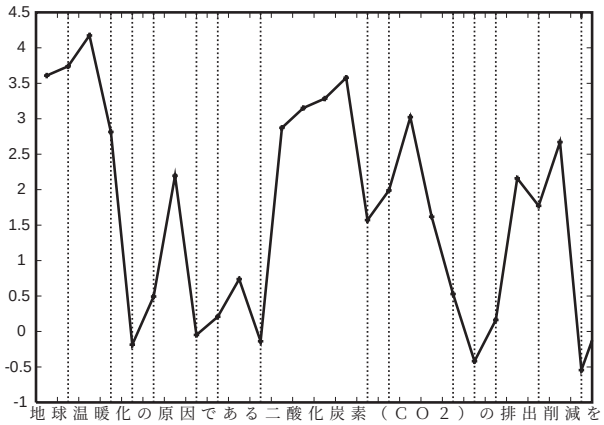

図 1: 解析対象の文に対する評価

3.3 形態素解析の誤り訂正

提案する手法の概要を図 2 にまとめる.

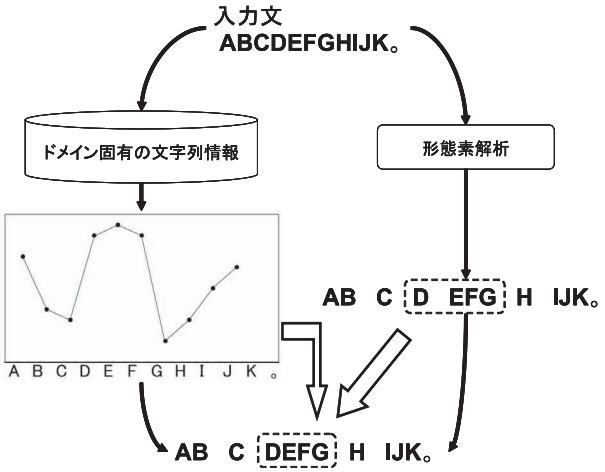

図 2: 提案手法の概要

具体的には，ドメイン固有の文字列の関する情報から，隣接スコアを算出し，このスコアが閾値以上の文字間で形態素解析ツールが分割を行っている場合，それを修正する.

4 実験

表 1 に実験の条件を示す.

表 1: 実験の条件

形態素解析器	chasen-2.29
形態素辞書	ipadic-2.5.1
コーパス	毎日新聞 99 年版
ドメイン	科学
文書数 (文数)	633(9098)
評価文	asahi.com
ドメイン	科学・自然

上記の条件で実験を行った結果，抽出することができた複合語，未知語の例を表 2 に示す.

表 2: 抽出できた単語の例

単語	chasen の出力
地球温暖化	地球 ^{N} + 温暖 ^{N} + 化 ^{N}
携帯電話	携帯 ^{N} + 電話 ^{N}
ガンマデルタ T 細胞	ガンマデルタ ^{U}+T^{U}+細胞 ^{N}
内視鏡	内 ^{N} + 視 ^{N} + 鏡 ^{N}
遡上	遡 ^{V} + 上 ^{N}
混獲	混 ^{U} + 獲 ^{V}

なお，N は名詞，U は未知語，V は動詞を表している.

「地球温暖化」などの例は，名詞の過分割を適切に検出している例と言える．また，未知語（「ガンマデルタ」等）を含む複合語に関しても適切に検出している.

湯本らは，専門用語の抽出手法を提案している [3]. しかし，この手法では，形態素解析済のテキスト中の名詞が連続する部分を対象としている．それに対して，提案手法においては,「遡上」,「混獲」の例のように，形態素解析器が名詞の一部を誤って動詞と判定した場合でも正確に抽出することが出来ており，形態素解析器の誤りに対してロバストであると言える.

また，表には示していないが,「気候変動枠組み条約第 3 回締約国会議」の様な d_{max} よりも長い複合語も抽出されている．これはドメイン固有の単語には特定のパターンがあることを示唆している.

5 まとめと今後の課題

提案手法は，処理を施していない生のコーパスに基づく字面の情報から，ドメイン固有の情報量を得て，辞書に基づく形態素解析の結果を修正することを試みている．二つの処理は完全に独立しており，また，生のコーパスを利用できることにより，効率のよい手法である，ということができる.

今後の課題として，以下の様な事が挙げられる.

- 適切な閾値の設定方法に関する検討.
- 文末表現のような頻出する表現について隣接スコアが高くなるため，適切な処理を施すようにする.
- D-bigram の文字間の距離にある程度の柔軟性を持たせることにより，新しい概念を抽出できるようにする．[1]

参考文献

[1] 内山将夫. “形態素解析結果から過分割を検出する統計的尺度”. 自然言語処理, Vol. 6, No. 7, pp. 3–25, 1999.

[2] 延澤志保, 佐藤健吾, 斎藤博昭. “ドメイン固有の文字列情報の組み込みによる形態素解析の精度の向上”. 自然言語処理研究報告, No. 9, pp. 21–40, July 2002.

[3] 湯本紘彰, 森辰則, 中川裕志. “出現頻度と連接頻度に基づく専門用語抽出”. 情報処理学会第 145 回自然言語処理研究会, 2001.

[1]「地球温暖化」の様な単語は「地球の温暖化」という表現から発生したと考えられる.