

オブジェクト列変換による単語分割方式

4M-04

加来 航[†]電気通信大学 電子工学科[‡]長井 隆行[‡] 樽松 明[‡]電気通信大学 電気通信学研究科[‡]

1. はじめに

日本語の文章は通常、英語などの言語の場合と異なり、テキストに単語境界を表記せず、連続した文字で記述される。しかし機械翻訳などにおいては、形態素解析において品詞同定の前段階に単語境界の推定が行われる。形態素解析においては、辞書を用いて文章を単語に分割していたが、未知語(辞書に無い単語)が出てくると正しい分割は難しい[1]。一般的に自然言語には、未知語が多数存在する。そのため予め用意した辞書による分割では、ニュースや特許情報、流行語や若者語で記述された情報には対応できない。

日本語形態素解析ソフト茶筌[3]では接続コスト最小法を用いているが、ここで使用される品詞接続コストと単語コストの具体的な値は実験的に決定しており、ある分野に調節したコストでは、他の分野では不適切なこともある。このように対象分野へのパラメータの適応や保守が難しいことが接続コスト最小法の問題点である。

また、単語の統計的な性質を用いた日本語単語分割法[3]、[4]も提案されているが、未知語に対しては十分ではない。

以上のような観点から本稿では、字面情報に着目し頻度情報に基づいてオブジェクトを自動的に生成するオブジェクト列変換という方法で、分野に依存せず、辞書の整備の不要な単語分割方法を提案し、実験結果を示す。

2. オブジェクト列変換

この変換は、任意のオブジェクト列の一部を内在する上位オブジェクトで表記し直す変換である。本稿では単語分割のための変換として使うので、文字列を単語列で表記し直す変換となる。アルゴリズムを次に示す。

簡単のため、変換対象を構成しているオブジェクトを object、上位オブジェクトを OBJECT と表記する。

- (1) 変換対象の object 列(Iとする)を、先頭から順に一定の大きさ(N[object]とする)の識別スコープを用いて、OBJECT 辞書(Dとする)を参照しながら、既知 OBJECT(Dに存在する object 列)と、未知領域(object 列)の2つに分ける。
- (2) 未知領域を定められた大きさの範囲(M[object]とする)で分割し、順に未知領域知識(Uとする)に記憶する。
- (3) 未知領域が定まるたびにUに記憶されているすべての未知領域と比較し、一致する部分(object 列)を抽出する。
- (4) 抽出した object 列を OBJECT の構成物情報として、Dに追加する。
- (5) (1)から(4)を繰り返す、Iの一部を既知 OBJECT に変換し出力する。

このように、自ら辞書を作りながら変換のための解析に使うことを特徴としている。

3. 単語分割システム

先に述べたオブジェクト列変換を使い、文章を単語に分割する方法を以下に示す。

- (1) 分割対象のテキスト文を文字の列(object 列)とみなし、すべての文字に識別記号を割り振る。
- (2) オブジェクト列変換をする。
- (3) 生成された辞書を用いて、文章を単語分割する。

このほか、予め初期知識として、句読点、助詞、記号などを登録することによって、一般の文章の分割性能の向上を図ることができる。

単語分割システムにおける処理の流れを図1に示す。

4. 実験

オブジェクト列変換により単語分割を行った例を示す。実験条件と結果は下記の通りである。

<条件>

初期知識:助詞、記号 計15語

分割対象:毎日新聞の記事6つ

識別スコープ長N:12[オブジェクト]

未知領域長M :12[オブジェクト]

Word Segmentation by Object Sequence Conversion

[†]Wataru Kaku, [‡]Takayuki Nagai, [‡]Akira Kurematsu

[†]The Univ. of Electro-Communication, Dept. of Electronic Engineering

[‡]The Univ. of Electro-Communication, Graduate School
Electro-Communication

<結果(一部)>

(前略)|ゲーム|業界|の|苦戦|が|続い|ている|。|今|月|二|日|、|ゲーム|ソフト|の|カプコン|が|一|九|九|四|年|度|の|業績|を|大幅|に|下方|修正|した|の|を|はじめ|、|一|部|には|セガ|・|エンタープライゼス|の|業績|の|悪化|も|伝え|られて|いる|。(中略)|今|年|末|の|商戦|から|来|年|に|かけ|、|ゲーム|機|、|ソフト|とも|次世代|ゲーム|の|時代|に|移行|する|転換期|。|急成長|を|遂げ|て|きた|ゲーム|業界|だが|、|生き残り|を|かけた|ソフト|開発競争|が|さら|に|激化|し|そう|だ|。|

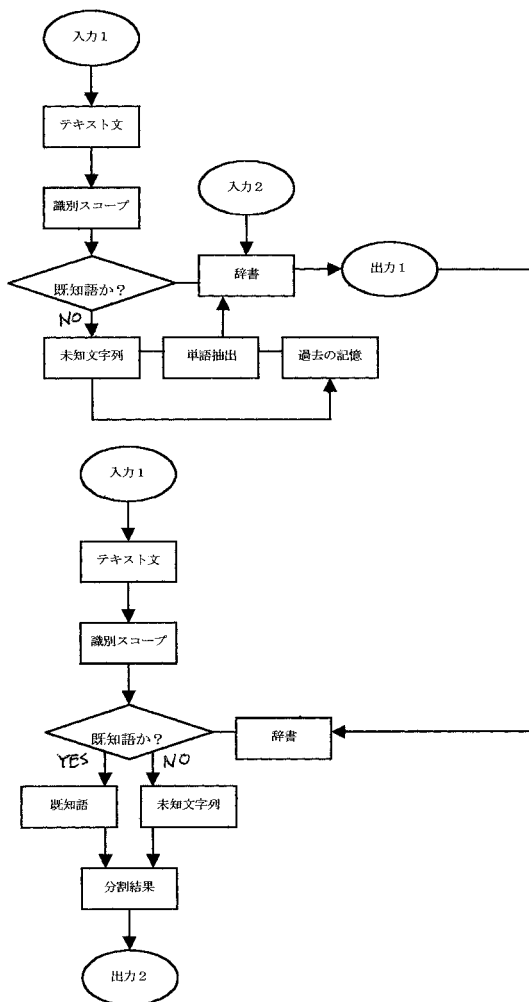


図1 単語分割システム

5. 結果

辞書を維持したまま1から6まで記事を順に入力して単語分割し人が単語分割した結果と比較した。評価項目は、

学習率＝正しい単語の数／全体の新出単語数、
学習精度＝正しい単語の数／学習した単語の総数、
分割精度＝ 正しく分割できた単語数／全体の単語数
を用いた。

ここで学習単語とは、システムが記事から学習し辞書に登録した単語である。また全単語数とは人が記事を単語分割した際の分割数である。実験結果を次の表1に示す。

表1 システムの性能

記事	[1]	[2]	[3]	[4]	[5]	[6]
文字数	307	349	1337	573	383	947
新出単語数	65	54	261	175	105	172
学習単語数	14	22	37	21	21	31
全単語数	150	178	723	306	195	516
学習率[%]	20	33	8	12	20	18
学習精度[%]	90	77	60	57	52	87
分割精度[%]	90	93	89	82	86	92

記事3、4、5で学習精度が低下しているのはシステムが数詞を知らなかったためであると考えられる。本稿で述べた方法では、平均して90%近い分割精度が保てていることがわかった。

6. まとめ

本稿では、オブジェクト列変換により、わずかな初期知識のみで、未知語の多い新聞記事を高精度に単語分割できることを示した。

今後の課題として、大量のデータを用いての方法の有効性を確かめる必要がある。また、本方法では単語の識別に右方向最長一致法を使用しており、この方法は識別が不完全であるので、他の方法と組み合わせる必要があると考えられる。また、解析量が増えるにつれて、誤った学習により性能が低下していくという問題があるので、単語の正誤確認機能を持たせる必要がある。

参考文献

- [1] 田中 穂積、“自然言語解析の基礎”、産業図書
- [2] 松本裕治、北内 啓、山下 達雄、平野 善隆、今一 修、今村 友明、日本語形態素解析システム「茶筌」version 1.0 使用説明書、奈良先端科学技術大学院大学(1997)
- [3] 伊東 伸泰、西村 雅史、“N-gram を用いた日本語テキストの単語単位への分割”、情報処理学会 自然言語処理 122-9、(1997)
- [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”、情報処理学会誌 Vol.40 No.9、(1999)