

## 帰納的学習を用いた単語分割手法の日本語 ・中国語・韓国語に対する有効性

王 忠建 荒木 健治 栃内 香次

北海道大学大学院工学研究科  
〒060-8628 札幌市北区北13条西8丁目

E-mail : {wzj,araki,tochinai}@media.eng.hokudai.ac.jp

インターネットの普及に伴い、様々な言語のテキスト情報が急増している。従って、情報検索、機械翻訳などの言語処理技術が多言語に対処できることが望まれる。これらの言語処理技術はすべて単語を基本単位として行われているため単語の間にスペースを入れる習慣のない言語ではまず文を単語に分割する必要がある。そこで、我々はこのような言語を対象とする単語分割手法を提案し、その有効性を既に確認した。本手法では帰納的学習を用いて文字列の共通部分、差異部分を再帰的に抽出し、単語として最も確からしいものを単語とし、文を単語に分割する。本稿では異なる言語を用いた実験結果から本手法の多言語に対する有効性を確認する。

キーワード： 単語分割，帰納的学習，多言語

## Effectiveness for Japanese, Chinese and Korea of Word Segmentation Using Inductive Learning

Zhongjian Wang Kenji Araki Koji Tochinai

Graduate School of Engineering, Hokkaido University  
N13-W8, Kita-ku, Sapporo 060-8628, JAPAN

E-mail : {wzj, araki,tochinai}@media.eng.hokudai.ac.jp

With the developing of the Internet and popularization of computers, a large amount of text information in different languages on the Internet are increasing explosively, so that it is necessary to develop a common method that can deal with multi-languages. We have proposed a method for non-segmented languages. In our method, we extract recursively common parts and different parts of a character string in text as word candidates by using inductive learning. We segment a sentence into words by using the certainty of extracted word candidates as a word. We only use the surface information of a text, so that the method is not dependent on each specific language. To confirm generality of proposed method, we carried out experiments with different languages.

Keywords: word segmentation, inductive learning, multi-language

## 1 はじめに

分かち書きをされていない言語，例えば，日本語，中国語，タイなどのような言語の自然言語処理において単語分割は欠かすことができない処理過程であり，機械翻訳，情報検索，及び音声認識などでは全て単語を基本的な単位として行われる．また，インターネットの普及によって様々な言語のテキスト情報が急増している．従って言語処理技術が多言語に対処できることが望まれる．これらの言語処理技術はすべて単語を基本単位として行われているため単語の間にスペースを入れる習慣のない言語ではまず文を単語に分割する必要がある．我々はこのような言語を対象として単語分割手法を提案し，その有効性を既に確認した [1][2]．

本手法では帰納的学習を用いて文書中に重複して出現する文字列の共通部分と差異部分を再帰的に抽出することにより，単語として確実性の高いものを単語候補として抽出する．また，抽出状況により抽出された単語候補を分類し，辞書に登録する．そして，登録された単語候補を単語としての確実性の高い順に用いて文を単語に分割する．本手法においては，表層レベルの字面情報のみから単語分割が行えるので言語に依存せず多言語に対応できるという利点がある．また，本手法は特定の言語，分野に依存せず，多言語へ適応できる．本手法の汎用性を評価するために行った実験では初期状態を一定とするために最初に辞書を仮定せず，空の状態で行った．本手法は文書中から共通部分，差異部分を抽出することにより分野ごと，ユーザごとに適応する単語辞書を生成し，文を単語に分割するので，どのような分野，ユーザにも適応が可能であるという特徴がある．

本稿では同一アルゴリズムで日本語，中国語，韓国語の文書を用いて本手法の分かち書きされていない言語に対する汎用性の評価実験を行った．日本語，中国語 [3] のような分かち書きされていない言語に対して韓国語では文節単位まで分かち書きされているが単語単位までは分かち書きされていない [4]．本稿は本手法の汎用性を確認するために韓国語のスペースを除いてべた書き文を生成し，ついで本手法により分かち書

きされていない言語のように単語に分割する実験を行った．実験の結果から本手法の多言語に対しての汎用性をもつことが確認された．

## 2 処理過程

本手法では帰納的学習を用いて入力文書において重複して出現する文字列を単語候補として再帰的に抽出する．また，抽出状況により抽出された単語候補 (WS) を単語とする確実性の高い順に分類し (WS: S2, S1, S3)，辞書に登録する．辞書に登録する情報は WS の出現頻度 FR (Frequency)，WS を用いた分割の正分割度数 CR (Correct Segmentation Frequency)，WS を用いた分割の誤分割度数 ER (Erroneous Segmentation Frequency)，WS の長さ LE (Length) がある．本手法のアルゴリズムの詳細について既に文献 [2] で述べているので，本稿では本手法の概要のみを述べる．

### 1. 既知語による分割:

辞書中に既に登録され，かつ単語として確認された CW と入力文中の部分文字列をマッチングすることにより分割候補のリストを作成する．分割は文の先頭から行う．分割候補が複数個存在する場合，正しい分割候補は式 (1) に示すゆう度評価関数，及び辞書に登録された語候補の誤分割度数，正分割度数，出現頻度，単語長，分割位置などの順に決定し，分割を行う．CW で分割できない場合，WS を S2, S1, S3 の順に用いて分割候補のリストを作成し，分割を行う．

$$LEF = FR + \alpha \times CS - \beta \times ES + \gamma \times LE(1)$$

式中の FR, CS, ES, LE それぞれは登録された CW あるいは WS の出現頻度，正分割度数，誤分割度数，長さである． $\alpha, \beta, \gamma$  は重み係数であり予備実験により決める．本稿では  $\alpha=1, \beta=70, \gamma=50$  を用いた [2]．

### 2. 未分割部分に対する語の推測:

既に登録された CW, WS を用いて分割されなかった部分文字列については帰納的学習を用いて未知語の推測を行う．文中に複数回現れる文字列は単語としての確実性が高いと考えられるので，これらの文字列

から再帰的に共通部分と差異部分を抽出することにより未知語を推測する．また文字列の抽出は再帰的に二段階すなわち共通部分の抽出と高次共通部分の抽出で行い，かつ，抽出される状況により単語とする確実性の高い順に三つのクラス S2, S1, S3 に分類される．

(a) 共通部分の抽出

二文字以上の文字列が文書中に重複して出現する場合，共通部分として抽出し S1 (Segment one) と呼ぶ．

(b) 高次共通部分の抽出

高次共通部分とは抽出された共通部分，差異部分から更に再帰的に共通部分，差異部分を抽出したものである．これは，一つの S1 が複数の単語によって構成される可能性があるからである．このような再帰的な共通部分，差異部分の抽出により，単語としてより確実性の高いものを抽出することができる．高次共通部分は単語として最も確実性が高いと考えられる．S1 同士の共通部分が存在，あるいは，S1 に別の S1 が含まれている場合，S1 同士の共通部分を抽出して，S2 (Segment two) とする．残りの差異部分は S3 (Segment three) とする．

3. 推測された語による分割:

既に登録されている CW, WS を用いても分割されなかった部分文字列を未知語の推測により得られた WS で分割を行う．分割は S2, S1, S3 の順で用いる．分割候補が複数個ある場合，1. 既知語による分割と同じ様に正しい分割候補を決定する．ここで 2 と 3 の処理は未知語の推測が出来なくなるまで繰り返して行う．

4. フィードバック処理:

ユーザが分割結果の正誤を判定し，分割誤りの校正，未分割部分の分割を行う．システムが校正した結果と分割結果とを比較することによりフィードバック処理を行い，辞書に登録された単語の頻度，正分割度数，誤分割度数などの情報を更新し，未分割部分の単語の登録を行う．

### 3 実験

本手法の異なる言語への汎用性及び異なる分野への適応性を確認するために，日本語，中国語，韓国語への適応実験を行った．実験結果を評価するために，式 (2), (3), (4) を用いた．本稿で正分割数とは分割結果において人間によって与えられた分割結果と完全に一致する数である．未分割数とは未分割文字列を，正しく分割した際に用いられる単語数である．誤分割数とは総単語数から正分割数と未分割数を除いた数である．

$$\text{正分割率} [\%] = \frac{\text{正分割数}}{\text{総単語数}} \times 100 \quad (2)$$

$$\text{誤分割率} [\%] = \frac{\text{誤分割数}}{\text{総単語数}} \times 100 \quad (3)$$

$$\text{未分割数} [\%] = \frac{\text{未分割数}}{\text{総単語数}} \times 100 \quad (4)$$

#### 3.1 実験データ

中国語の実験データは Sinica コーパス [5] から建築学の文書 145,727 単語，経済学の文書 113,000 単語と電子工学の文書 116,110 単語を用いた．三つの分野の文書の構成については建築学には建築美学，建築評論及び建築新聞などがある．経済学には経済システム，経済政策及び経済理論などの文書がある．電子工学には電子学，通信工程及び機械工程，核工業などの文書がある．合わせて 374,837 単語の文書をデータとして分野ごとに対する適応性の評価実験を行った．

日本語の実験データは EDR コーパス [6] から 389,230 単語の文書を用いた．また，日本語の文書は分野ごとに整理されずランダムであった．また，参考のために韓国語の実験を行った．韓国語の実験データは Web から引用した 91,033 単語の小説の文書をスペースを除いて分かち書きされていない言語に変換して同様の実験を行った．三つの言語の文書を用いて本手法の言語ごとに対する汎用性の評価実験を行った．

#### 3.2 実験手順

実験は同一アルゴリズムを用いて辞書が空の状態から始め，三つの言語の文書を一言語づつ，

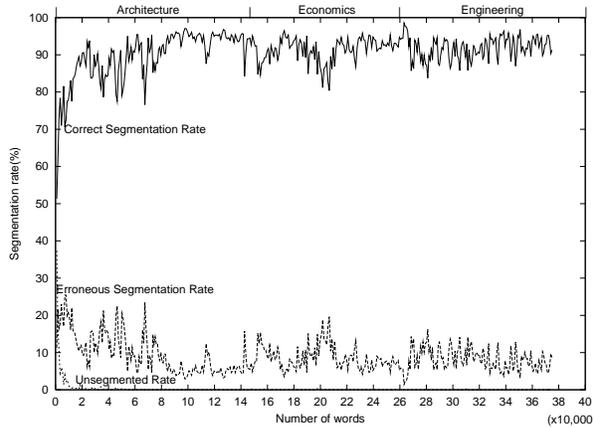


図 1: 中国語の 3 分野の文書の分割率の推移

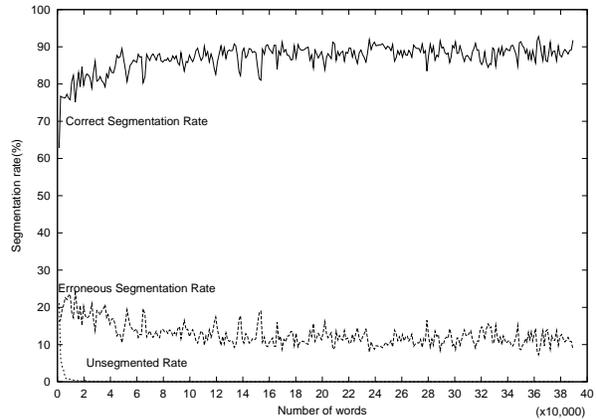


図 2: 日本語の文書の分割率の推移

一つの言語の文書に対して約 100 単語の段落を一段落ずつ入力して行った。フィードバック処理ではユーザが分割結果の誤りを校正し、システムが分割結果と校正済み結果とを比較することにより、辞書に登録された単語候補情報を更新し、未登録語の登録を行う。

### 3.3 実験結果

三つの言語の実験結果を表 1 と図 1, 図 2, 図 3 に示す。図 1, 図 2, 図 3 は中国語文書, 日本語文書, 韓国語文書それぞれの正分割率, 誤分割率, 未分割率の推移を表している。表 1 は日本語の平均正分割率, 韓国語の平均正分割率, 中国語の 3 分野のデータを順番に入力した場合のそれぞれの分野の平均正分割率と全体の総平均正分割率を示している。

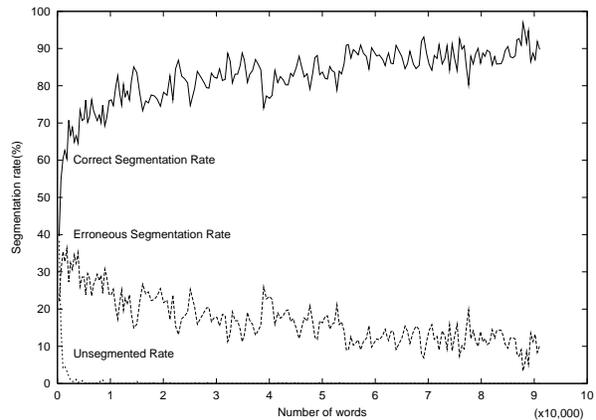


図 3: 韓国語の文書の分割率の推移

## 4 実験の考察

### 4.1 有効性

表 1 に示されるように中国語 3 分野の文書それぞれの平均正分割率は 89.3%, 91.2% と 91.9% で, 総平均正分割率は 90.6% となった。上記の 374,837 単語の文書中には人名, 地名などの固有名詞, 専門用語などが含まれているが, それに対して特別な処理を行わなくても帰納的学習を用いて未知語を推測できることが確認された。最初は辞書が空なので未知語を推測しながら分割を行っている。未知語が推測され登録される

単語数が増加するにつれて正分割率が向上している。約 23,000 単語が処理されたとき, 正分割率が約 95% に達したが, その後辞書に登録された CW, WS が増加するのに伴い, 曖昧分割が原因で誤分割率が大きくなっている。曖昧分割とは同一文字列を単語に分割するとき複数の分割可能性があるということである。しかし, 約 80,000 単語が処理されたとき, フィードバックの効果により登録された CW, WS の正分割度数, 誤分割度数などが更新され, 誤分割率が低下し, 正分割率が上昇している。分野が変化したとき, 正分割率は一時的に低下している。これは専門用語などの未知語が出現したことが原因であるが, 帰納的学習で未知語を推測し, 獲得することにより正分割率は再び上昇している。

表 1: 中国語, 日本語, 韓国語の分割実験結果

言語	中国語 (総単語数: 374,837)				日本語	韓国語
	建築学	経済学	電子工学	平均	ランダム	小説
単語数	145,727	113,000	116,110	124,945	389,230	91,033
CSR[%]	89.3	91.2	91.9	90.6	87.5	80.8
ESR[%]	10.1	8.4	8.1	9.0	12.3	18.9
USR[%]	0.6	0.4	0.0	0.4	0.2	0.3

また, 局所的に変化する個所が見られ, その箇所の正分割率が低下しているが, これは分野内での細かい変化のために未知語が出現したことが原因と考えられる. 例えば, 建築学には建築美学, 建築評論及び建築新聞などの文書があるため, 新しい用語が出現し, 正分割率に影響を与える.

図3に日本語文書の単語分割の正分割率, 誤分割率, 未分割率の推移を表している. EDR コーパスのテキスト文書は分野ごとにまとめられていないため, 図2の分野ごとの中国語の実験結果と比べると, 正分割率の大きい変化が見られなかった. しかし, これは異なる分野の文書が混在しているため専門用語などが徐々に発生し, 用語の大きな変化がないことが原因と考えられる. また, 処理された単語数が増加するに従って正分割率が徐々に上昇している. 実験の終わりごろ, 正分割率はほぼ90%に達したが平均総正分割率は87.5%であった. 同一アルゴリズムで異なる言語を用いた実験結果より本手法は言語ごとに適応すること, 即ち, 汎用性を持つことが確認された.

#### 4.2 学習機能の考察

本手法で用いた帰納的学習の有効性について以下の2通りの評価実験を行った。「実験1. 本システムのアルゴリズムで帰納的学習を用いてWSを抽出することにより未知語を推測し, 分割を行った実験」「実験2. WSを抽出せず, 文を入力して, 分割結果の誤りを校正して分割を行った実験」.

実験には中国語の建築学分野のデータと経済

学分野のデータ, 総単語数約180,000単語の文書を用いた. 実験1の正分割率と実験2の正分割率の推移を図4に示す. 実験1の実験2に対する正分割率の改善率の推移を図5に示す. 改善率を式(2)に示す. 正分割率1と正分割率2はそれぞれ実験1, 実験2の正分割率である. 正分割率の推移から, 本手法の学習の効果が明らかに表れた. 実験1の場合, 実験は辞書が空の状態から急速に安定した正分割率になったことが分かる. また, 分野が変化するとき, WSの抽出がある場合とない場合の相違が分かる. 分野が変化するとき, 実験1ではシステムが急速に分野に適応し, 正分割率がより早く上昇することが表れている. 図5における実験の最初と分野が変化するときの正分割率の改善率の大幅な増加からも本手法の学習機能の有効性が示されている.

改善率が局所的に振動する個所があるが, それは分野内での文書の細かい変化によって未知語が出現したことが原因と考えられる. 例えば, 建築学には建築美学, 建築評論及び建築新聞などの文書があるため, 新しい単語が出現し, 正分割率に影響を与える. そのような局所の振動がある部分に対する本手法の学習機能の効果が, 図5の改善率の変化により示されている. 実験の最初の20,000単語を処理したときの正分割率の平均改善率は20.4%である. システムが処理した単語数の増加につれて学習が進んでいるので, 共通部分と差異部分を抽出することにより未登録語の推測が大幅に減少し, 改善率が下がっている. また, 分野が変化するとき実験1と実験2とを比べると実験1の方が急速に分野に適応できることが図4により明らかに示されて

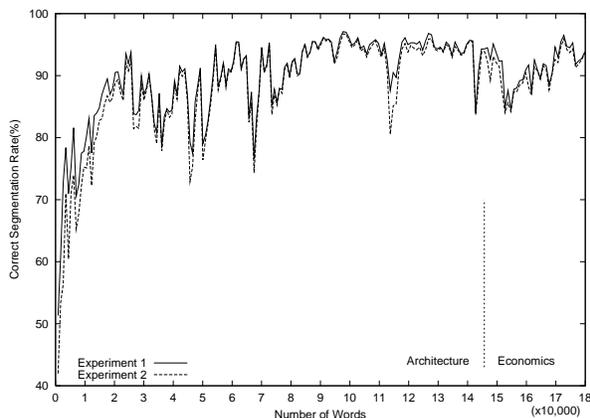


図 4: 実験 1 と実験 2 の正分割率の推移

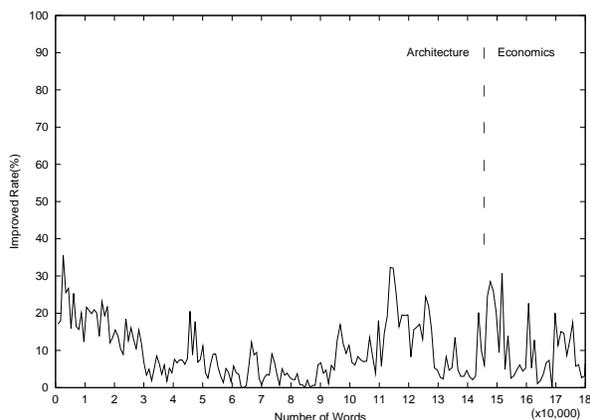


図 5: 改善率の推移

いる．よって再帰的に共通部分，差異部分を抽出する帰納的学習は汎用的単語分割手法として有効であると考えられる．

$$\text{改善率} [\%] = \frac{\text{正分割率} 1 - \text{正分割率} 2}{100 - \text{正分割率} 2} \times 100 \quad (5)$$

### 4.3 誤分割の考察

中国語の実験結果から実験の始めの 1,000 単語と実験終わりごろの 1,000 単語を用いて誤分割の内訳を考察した．最初の 1,000 単語の誤分割の 15% のうちの未登録語による誤分割率は 88% を占めている．また，実験の終わりに近い 1,000 単

語の誤分割の 7.2% のうちの曖昧分割による誤分割は 83% を占めていることが分かった．これは，実験の最初では未登録語が多いため誤分割が多数発生し，実験が進んで辞書に登録された単語が増えるのに伴い，未登録語による誤りの割合が下がると同時に曖昧分割による誤りの割合が大きくなったものと考えられる．

## 5 おわりに

本稿では帰納的学習を用いた単語分割手法の多言語に対する有効性を確認するために三つの言語，日本語，中国語及び韓国語を用いた単語分割実験を行った．中国語の 3 分野の文書に対して分野ごとの適応性，及び日本語と中国語の実験結果を比較することにより，本手法は特定の言語に依存することなく複数の言語に動的に適応できることが確認された．今後の課題としては曖昧分割の解消と帰納的学習を用いて品詞を付与することを計画している．

## 参考文献

- [1] 荒木健治, 柝内香次, “ 帰納的学習による語の獲得および確実性を用いた語の認識, ”電子情報通信学会論文誌, D-II Vol.J75-D-II, No.7, pp.1213-1221, July 1992.
- [2] 王 忠建, 荒木健治, 柝内香次, “ 帰納的学習を用いた単語分割手法の多言語に対しての汎用性の評価, ”電子情報通信学会技術研究報告, TL2001-44, NLC2000-79 (2001-03).
- [3] 香坂順一, 中国語の単語の話, 光生館, 東京, 1971.
- [4] 鄭明石, 鄭喜盛, 現代韓国語 基本文型と構造, 高麗書林, 東京, 1982.
- [5] “中央研究院平衡語料庫的内容與説明, ” Technical Report no.95-02, 中文詞知識庫小組, 1995.
- [6] 日本電子化辞書研究所, “EDR 電子化辞書使用説明書, ” 1995.