社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
TL99-43 (2000-03)

帰納的学習を用いた中国語の単語分割手法の性能評価

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

あらまし 本稿では帰納的学習を用いて文書中に重複して出現する共通文字列を共通パターンとして抽出し、学習することにより単語として確実性の高いものを用いて文を単語に分割する手法の中国語文への適用について述べる。本手法は文法、言語の知識などを用いないので多言語を処理できるという利点がある。本手法では辞書、分割ルールなどを予め用意する必要がなく、入力文中の未知語を推測して辞書を生成し、分割結果を校正した情報を用いることにより分割精度が上昇する。既に、日本語に対しての有効性が確認されているので、本研究では中国語に対する本手法の有効性を確認するための実験を行なっている。実験結果により本手法の汎用性が確認された。

キーワード:帰納的学習、中国語、単語分割、共通パターン、多言語

Evaluation of Chinese Word Segmentation Method 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

Abstract In this paper, we propose a word segmentation method for Chinese sentence using a common character string that occurs frequently in text, and call it common pattern. We considered that the common pattern has high probability as word. It is extracted as word candidate and registered in dictionary. Furthermore, in our method it is not necessary to prepare a dictionary and any segmentation rules for dealing with an ambiguous segmentation beforehand. We have confirmed that the proposed method is effective to Japanese word segmentation. In this paper, we describe the results of experiment for Chinese word segmentation using our proposed method. The results show that the proposed method is possible to use for multi-language word segmentation.

Key words: inductive learning, Chinese, word segmentation, common pattern, multi-language

1 はじめに

単語分割は自然言語処理における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，及び音声認識などは全て単語を基本的な単位として行われる。それゆえアジアの言語，例えば日本語，中国語，タイ語などのように単語分かち書きをしない言語を計算機で扱う場合，まず文書を単語に分割しなければならないという問題がある。我々はこの問題に対して多言語の単語分割処理を目指して帰納的学習アルゴリズムを提案し，日本語処理において有効性が確認されている [1]。本論文はそれを中国語処理に適用し，その有効性を実験により確かめるものである。

中国語の表記法は漢字を羅列するだけで，漢字の大部分が表意文字である。一個の漢字は一個の概念を持ち，一個の音節を持っている。単語は一個の漢字から六個の漢字 (もっと多い場合もある) で構成される。中国語は語形変化のない言語であり，動詞は形容詞としても，名詞としても使えることが多い。中国語では日本語の助詞に相当する語がないので，複合語の場合は文とみなすこともでき，主語，述語などの成分が存在する。文脈によって単語の切目と品詞の分類が異なる [2]。

例えば，“一把鉄鋸” の “鉄鋸” は名詞であるが，“鋸木頭” の “鋸” は動詞である。さらに，“鉄鋸” の “鉄” と “鋸” はそれぞれ形容詞と名詞に分けることもできる。単語を構成する同じ漢字が他の単語に出現し，更に一つの単語をとして存在し得る。このように，中国語の単語分割では未知語の認識と複数の分割可能性を持つ多義分割が主な問題である。以上述べた中国語の特徴と，さらに，単語の多義性及び未知語の発生のために，中国語の単語分割は非常に困難である。

中国語の単語分割に関する従来の研究はいくつか挙げられるが，大別して規則に基づく手法 [3][4][5]，統計的な手法 [6][7]，辞書規則と統計情報を結合する方法 [8][9] がある。文献 [4] は多義分割 (複数の分割可能性がある分割) が生じる種類と原因を分析し，まとめた規則を用いて多義分割を処理する。規則に基づく方法は予め単語辞書，多義分割を処理する規則を用意する。規則に基づく方法の正解率は辞書の大きさと規則の量に依存する。更に規則の抽出，整理及び更新には膨大な労力がかかる。文献 [8] には中国語の単語分割に対し，単語の出現確率を重みとする重み付き有限状態変換器を用いる方法が提案されている。文献 [9] は規則辞書とコーパスに基づく統計的な手法を用いて新聞の文書中の人名を認識する方法を提案している。統計的な手法では漢字の隣接情報を利用して単語の境界を決定している。一般に高精度な統計モデルを構成するためには大規模なタグ付けデータが必要である。統計的な方法は未知語の認識に対して比較的有効な方法である。また，辞書規則と統計的な情報を結合する方法では規則辞書に基づいて統計的な情報で規則が適用するかどうかを判断して，最適分割パスを決定している。

これらの方法は予め単語辞書，タグ付けコーパスなどの用意が必要である。しかし，すべての単語と各種規則を登録するのは不可能であり，大規模なタグ付けコーパスの作成にも多くの労力が必要となる。最近の研究では辞書，タグ付けコーパスを用いないで生コーパスから漢字間の隣接の統計的な情報を利用して単語分割方法も提案されている [10]。

我々の提案する手法はそれらの方法と異なり，予め単語辞書，規則辞書，タグ付けコーパスなどを準備する必要は一切ない。さらに，表層レベルの情報のみから単語分割が行えるので言語に依存しないという利点がある。本手法は字面の情報を利用し，文書に頻繁に出現する文字列を共通パターンとして抽出する。共通パターンを抽出することによって単語を獲得し，分割結果と校正済みの分割結果を比較することにより登録された共通パターンを単語とする優先順位を決め，文を単語に分割する。

2 概要

本手法の処理の流れを図 1 に示す。

まず，文書を入力して，すでに辞書に登録されている共通パターンを単語として用いて文書を単語に分割する。登録された共通パターンで分割されなかった部分を帰納的学習を用いて未知語を推測する。未知語の推測は文書中に複数回出現する文字列を共通パターンとして抽出する。抽出された共通パターンは抽出する位置，及び抽出状況によって，単語としての優先順位を決めて，辞書に登録する。そして，推測された語で分割を行う。そしてユーザが分割結果の正誤を判断し，誤りを校正して，その情報をフィードバックする。さらに，校正された結果より正しい単語を辞書に登録する。また，正分割結果の共通パターンの尤度を増やして，誤り分割結果の共通パターンの尤度を減らして，尤度を更新する。

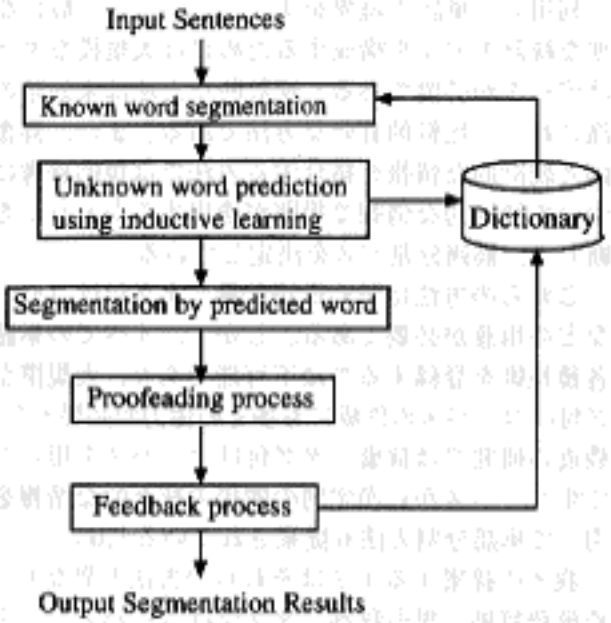

図 1: 処理の流れ

3 処理過程

3.1 既知語の分割

それまでに帰納的学習により獲得された単語を用いて文を単語に分割する。この分割過程は辞書に登録された共通パターンを単語としての尤度の高い順に行う。処理の手順を以下に示す。

(1) 入力文中の部分文字列と辞書中の共通パターンを比較する。

(2) 部分文字列と共通パターン全て一致する場合，この部分文字列を単語として分割を行う。

(3) 分割に複数の可能性がある場合，正分割候補の尤度関数の値の大きい順に分割を行う。尤度関数の値が同じ場合，誤分割度数の最小，正分割度数の最大，長さの最長と分割の位置の最前などの順に情報を用いて正しい分割候補を決定して分割を行う。尤度関数は以下の式で与えられる。

$$LEF(Likelihood\ Evaluation\ Function) = \alpha \cdot CS + FR - \beta \cdot ES + \gamma \cdot LE \quad (1)$$

ここで，LEF は尤度関数の値である。CSR (Correcr Segmentation Rate) は正分割度数，FR (Freqency) は単語の出現頻度，ESR (Erroneous Segmentation Rate) は誤分割度数，及び LE (Length) は単語の長さを表している。α と β，及び γはそれぞれ重み係数である。

(4) 複数の分割可能性の尤度関数の値が同じ場合，正分割候補は正分割度数の高いもの，誤分割度数の少ないもの，出現頻度の高いもの，分割位置が一番前のものと長さの大きいもの順に分割を行う。

3.2 未知語の推測

既知語で分割されなかった部分は帰納的学習を用いて未知語を推測する。未知語の推測は文中に表れる共通の部分文字列を共通パターンといい，この共通パターンを抽出することにより未知語を推測する。共通パターンは単語としての確実性が高いと考えられる。そこで共通パターンを抽出し，単語候補として辞書に登録する。また抽出された共通パターンは抽出された状況により単語とする確実性の高い順で分類される。図 2 に中国語の文書の例を示す。次にこの図を用いて未知語の推測を説明する。

中国民営科技企業経過多年的発展，目前正歩入高速発展的時期。民営科技活動，已経覆蓋了国民経済主要行業，成為中国発展高科技産業。

図 2: 中国語における未知語の推測の例

共通パターンの抽出手順を以下に示す。

3.2.1 共通パターンの抽出

この段階で抽出された共通パターンを S 1(Segment One) と呼ぶ。抽出の条件を以下に示す。

(1) 2 文字以上の長さの文字列が重複して出現する場合，共通パターンとして抽出する。図 2 の例において，抽出した S1 の共通パターンは" 中国"," 民営科技"," 発展"，及び" 国民" である。

(2) 他の共通パターンに含まれない出現位置が少なくともある。例えば," 科技" は抽出され，S1 とする。

3.2.2 高次共通パターンの抽出

抽出された S1 には他の共通パターンを含む場合には抽出することができる。この段階は，S1 からの共通パターンの再抽出を高次パターンの抽出という。高次共通パターンは単語として最も確実性が高いと考えられる。高次共通パターンの抽出条件を以下に示す。

(1)S2 と S3 の抽出　抽出された S1 に含まれている共通部分が存在する。あるいは，S1 に別の共通部分が含まれていて，さらに S1 に含まれない出現位置が少なくとも一つ存在する。この部分を抽出して，S2(Segment Two) と呼ぶ。残りの部分は S3(Segment Three) と言う。例えば，" 科技 (S1)" は" 民営科技 (S1)" に含まれ，" 民営科技" から" 科技" を抽出できる。" 科技" を抽出して S2 に所属させ，残りの" 民営" は S3 に所属させる。

(2) 一文字パターンの処理　一文字が分割済みの単語で囲まれている場合，この一文字を単語として抽出する。抽出した一文字のパターンは S2 といい，単語としての確実性が高いと思われる。図 1 において，" 高" は" 発展高科技" に" 発展" と" 科技" で囲まれ，両側は分割済みのため，" 高" を抽出して S2 に所属させる。それぞれの抽出した共通パターンを単語として確実性の高い順に辞書に登録する。

3.3　辞書の構造

帰納的学習で推測された単語候補を S1，S2 と S3 に分類して，さらに優先順位を付けて辞書に登録する。ついに，推測された共通パターンを用いて既知語で分割されなかった文を単語に分割する。表 1 には辞書の構造を示す。

表 1: The Construction of The Dictionary

Word	FR	CSR	ESR	LE	CL
中國	10	8	0	4	CW
發展	8	6	1	4	S1
國民	7	5	1	4	S1
科技	12	12	0	4	S2
高	21	14	4	2	S2
民營	6	0	2	4	S3

ここで，CL は登録された共通パターンの分類（Classification）である。CW(Correct Word) はフィードバック処理で確認された正しい単語である。

3.4　推測された語を用いた分割

既知語を用いて分割されなかった部分を推測された共通パターンを用いて分割する。文中の文字列は用いた共通パターンと同じの場合，分割を行う。これらの共通パターンを単語として用いるのが以下の順に従う。

(1)　CW に所属する単語を用いて分割する。CW に所属する単語はフィードバックで確認された正しい単語である。複数の分割候補が存在する場合，尤度関数の値が大きいものを優先に用いて分割する。

(1)　S2 に所属する共通パターンを用いて分割する。CW の単語がない場合，S2 の共通パターンを用いる。複数の分割候補が存在する場合，尤度関数の値が大きいものを優先に用いて分割する。

(1)　S3 に所属する共通パターンを用いて分割する。S2 の単語がない場合，S3 の共通パターンを用いる。複数の分割候補が存在する場合，尤度関数の値が大きいものを優先に用いて分割する。

(1)　S1 に所属する共通パターンを用いて分割する。S3 の単語がない場合，S1 の共通パターンを用いる。複数の分割候補が存在する場合，尤度関数の値が大きいものを優先に用いて分割する。

3.5　フィードバック処理

ユーザが分割結果を見て必要があれば，誤り結果を直す。そして誤りを含む分割結果と直した正しい分割結果をフィードバックする。システムはフィードバックされた正誤結果を比較しながら登録された単語，共通パターンの正分割度，誤分割度及び頻度，所属を更新する。正分割の場合，単語あるいは共通パターンの正分割度数を増加される。誤った分割の場合，誤分割度数を増加される。共通パターンの尤度を更新する方法は以下である。

/中//國民//營//科技//企業/經過/多//年/
/的//發展//，//目前//正//步入/高速//
發展//的//時期//。//民營//科技//活動//
，//已經//覆蓋//了//國民//經濟//主要//
行業//，//成為//中國//發展//高//科技//
産業//的//生力軍//。/

図 3: 誤りを含む単語分割結果

/中國//民營//科技//企業//經過//多//年/
/的//發展//，//目前//正//步入//高速//
發展//的//時期//。//民營//科技//活動//
，//已經//覆蓋//了//國民//經濟//主要//
行業//，//成為//中國//發展//高//科技//
産業//的//生力軍//。/

図 4: 直された単語分割結果

（1） 正分割結果の尤度の更新

正分割の場合，正分割に用いた単語の頻度と正分割度数を増加する。もしその単語の分類はCWに所属していない場合，CWに所属する。

（2） 誤分割結果の尤度の更新

（2－1） 正しい単語が辞書にない場合，正しい単語が頻度を1にして，分類CLをCWに所属して辞書に登録する。

（2－2） 正しい単語を辞書にある場合，システムは単語の頻度FRを1増加し，もし単語の分類がCWに所属していないなら，CWに所属する。

（2－3） 誤分割の原因は誤り単語を用いたとする時，誤り単語の誤分割度数を1増加される。

（3） 未分割結果の処理

（3－1） 正しい単語が辞書にある場合，システムは単語の頻度を1増やし，もし単語の分類CLをCWに所属していないなら，CWに所属する。

（3－2） 正しい単語が辞書にいない場合，システムは正しい単語を辞書に登録する。登録された単語の頻度を1にして，分類CLをCWに所属する。

図3，4は単語分割結果と直された結果の例を示す。フィードバックで登録された単語候補の尤度の更新過程を以下に例で説明する。

/中/，/国民/，/営/と"経過"はそれぞれ誤分割と未分割である。他の部分は正しい分割である。

正分割結果の処理： "科技"，"企業"などの正しい分割を用いた単語の頻度を1増やし，分類CLはCWに所属していないなら，CWに所属する。

誤分割結果の処理： "中"，"国民"，"営"などの誤分割を用いた単語の誤分割度数を1増加する。さらに，正しい単語が辞書にあれば，頻度を1増やし，分類CLをCWに所属する。正しい単語が辞書になければ，頻度を1にして，分類CLをCWにして辞書に登録する。

未分割結果の処理： "経過"は未分割部分である。もし"経過"が辞書にないなら，頻度を1，分類CLをCWにして登録する。もし"経過"が辞書にあるなら，頻度を1増やし，分類CLをCWに所属する。

4 評価実験

4.1 予備実験

まず，尤度関数の係数の最適値を決定するために予備実験を行った。200文の中国語文書（約9,195単語）を用いて尤度関数の係数を変化させ最適な正分割率が得られる係数を求めた。実験の結果を表2に示す。この実験結果から最適な係数はα=1，β=50，及びγ=10となった。

4.2 実験データ

実験のデータとしてSinica Corpus[11]から物理学と医学の二つの分野のデータを用いた。物理学のデータには物理学総論，力学，熱学，光学，現代物理学などの文書を含み，6万単語の文書である。医学のデータには基礎医学，生理学，病理学，内外科医学，精神病学などの文書を含み，7万単語の文書である。

4.3 実験結果

実験は辞書が空の状態から始め，二つの分野のデータを1,000単語ずつ入力して実験を行った。結果を表3と図5,6,7に示す。図5,6,7はそれぞれ正分割率，誤分割率と未分割率の推移を表している。

5 考察

実験結果を評価するために式(2),(3),(4)に示す正分割率，誤分割率，未分割率を用いた。実験は辞書が空の状態から行い，正分割率は平均90.9%となった。この13万単語の文書中には人名，地名などの固有名詞，専門用語などを含み，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分かる。

最初は辞書が空なので未知語の推測しながら分割を行い，未知語が推測されるのに従って正分割率が向上している。しかし，約5,000単語が入力された時，辞書に登録された単語候補が増加することに伴い，誤分割率が大きくなっている。約18,000単語が入力された時，フィードバックの効果により登録された共通パターンの正分割度数，誤分割度数などが更新され，誤分割率が低下している。

分野が変わったとき，正分割率は低下している。そ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。

また，局所的な変動が見られ，例えば図5のA,B,C点の正分割率が低下しているが，これは分野の細か

表 2: Preliminary Experiments of Optimum Coefficient

α	1	1	1	0	1	5	10	1	1	1	1
β	1	1	1	1	1	1	1	50	60	70	80
γ	0	10	20	10	10	10	10	10	10	10	10
CSR[%]	85.0	86.8	86.8	86.0	86.8	84.1	82.0	87.7	87.6	87.7	87.4
ESR[%]	12.0	10.2	10.1	31.7	10.2	12.6	14.8	9.4	9.4	9.4	9.6
USR[%]	3.1	3.0	3.1	3.1	3.0	3.3	3.1	3.0	2.9	3.0	3.0

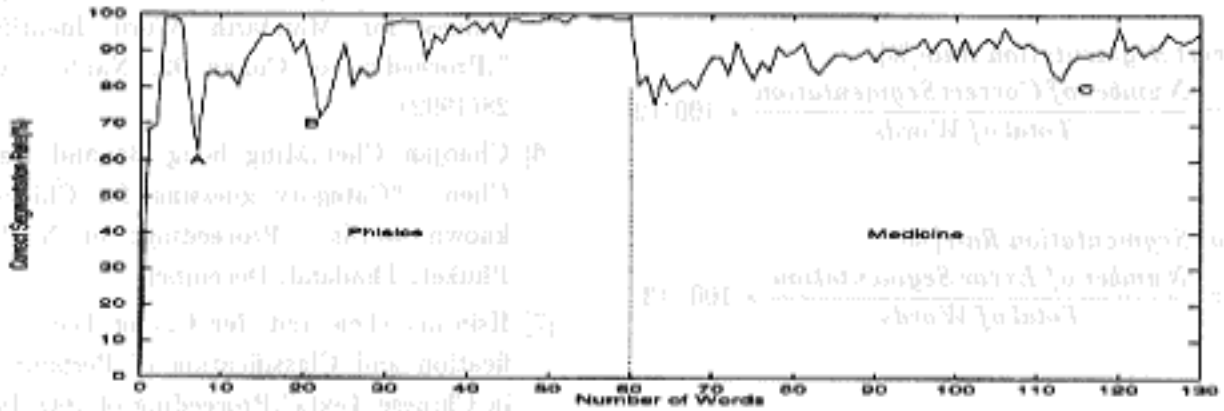

図 5: Change in Correctness Segmentation Rate

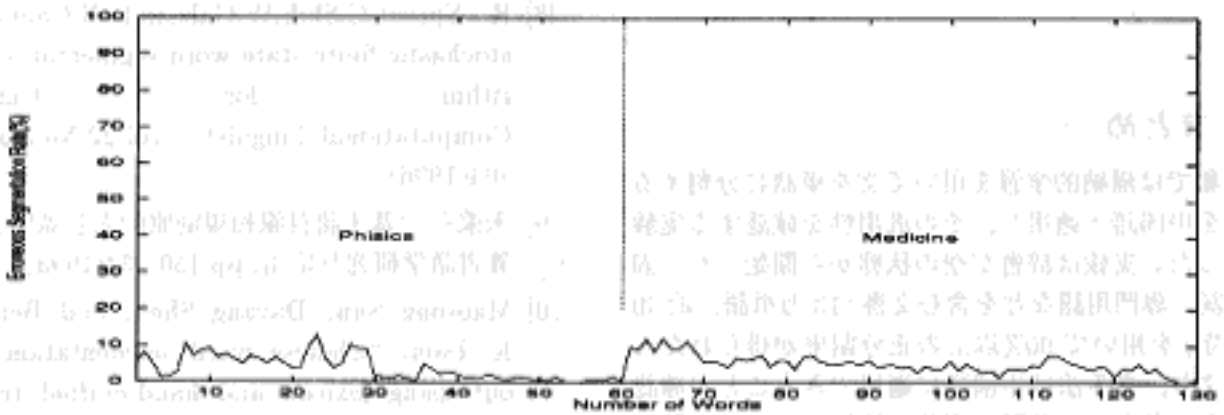

図 6: Change in Error Segmentation Rate

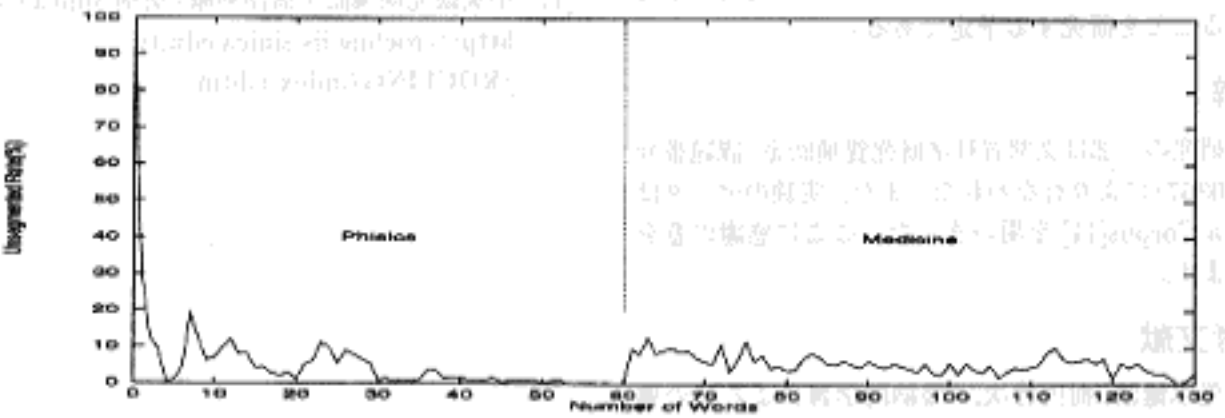

図 7: Change in Unsegmented Rate

表 3: Results of Experiment

CSR[%]	ESR[%]	USR[%]
90.9	5.9	3.2

い変化のために未知語が出現することが原因と考えられる。

$$Correct\ Segmentation\ Rate[\%] = \frac{Number\ of\ Correct\ Segmentation}{Total\ of\ Words} \times 100 \quad (2)$$

$$Error\ Segmentation\ Rate[\%] = \frac{Number\ of\ Error\ Segmentation}{Total\ of\ Words} \times 100 \quad (3)$$

$$Unsegmentation\ Rate[\%] = \frac{Number\ of\ Unsegmentation}{Total\ of\ Words} \times 100 \quad (4)$$

6 まとめ

本稿では帰納的学習を用いて文を単語に分割する手法を中国語へ適用し，その汎用性を確認する実験を行った。実験は辞書が空の状態から開始した。固有名詞，専門用語などを含む文書 (13 万単語，約 40 万文字) を用いて 90%以上の正分割率が得られたことにより，本手法は中国語に適用できることが確認された。さらに，分野の変化に追従し，どの分野に適応できることも確認された。

今後，実験の規模をもっと拡大し，品詞情報を付与することを研究する予定である。

謝辞

本研究の一部は文部省科学研究費補助金 (課題番号 10680367) により行なわれた。また，実験のデータは Sinica Corpus[11] を用いました。ここに感謝の意を表します。

参考文献

[1] 荒木健治, 桃内香次, "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 No7 pp.1213-1221(1992-7)

[2] 香坂順一, "中国語の単語の話", 光生館 (1971)

[3] 呉勝遠, "一種漢語分詞方法", 計算機研究与発展, Vol.33 No.4 pp.306-311(1996-4)

[4] 顧湃ら, "漢語自動分詞的近似匹配算法及其在 QHFY 漢英機器翻訳系統中的実現", 計算言語学研究与応用, pp.132-138(1993)

[5] Liang-Jyh Wang, Wei-Chuan Li and Chao Huang Chang: "Recognizing Unregistered Names for Mandarin Word Identification ",Proceeding of Coling 92, Nantes, Aug.23-28(1992)

[6] Chaojan Chen,Ming hong Bai,and Keh jian Chen: "Category guessing for Chinese unknown words", Proceedings of NLPRS'97, Phuket, Thailand, December

[7] Hsin-his chen and Jen-Chang Lee: "Identification and Classification of Perpoor Nouns in Chinese Texts",Proceeding of 16th International Conference on Computational Linguistics. Copenhagen,Denmark,Augst 5-9(1996)

[8] R. Sproat,C.Shih,W.Gale,and N.Chang, "A stochastic finite-state word-segmentation algorithm for Chinese", Computational Linguistics,vol.22,No.3,pp.377-404(1996)

[9] 宋柔ら, "基于語料庫和規則庫的人名識別法", 計算言語学研究与応用, pp.150-154(1993)

[10] Maosong Sun, Dayang Shen, and Benjamin K Tsou. "Chinese word segmentation without using lexicon and hand-crafted training data".17th International Conference on Computational; Linguistics, pp.1265-1271(1998)

[11] 中央研究院漢語平衡語料庫 (簡稱 Sinica Corpus) http://rocling.iis.sinica.edu.tw /ROCLING/index_c.htm