論　文

帰納的学習を用いた単語分割手法の中国語への適用とその性能評価

王　忠建† 荒木　健治† 栃内　香次†

Application for Chinese and Performance Evaluation of Word Segmentation Method Using Inductive Learning

Zhongjian WANG†, Kenji ARAKI†, and Koji TOCHINAI†

あらまし　我々は多言語を処理できる単語分割手法の開発を目指して帰納的学習により文書を単語に分割する手法を提案している．本手法においては文書中に重複して出現する文字列の共通部分及び差異部分を再帰的に抽出し，学習する．そして，単語として確実性の高いものから順に用いて文を単語に分割する．よって本手法では辞書，分割ルールなどをあらかじめ用意する必要がなく，入力文中に重複して出現する文字列の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し，辞書に登録する．更に分割結果を校正した情報を用いることにより文書を単語に分割する能力が向上する．このように本手法は固有の言語に依存した知識を用いないので，多言語を処理できるという利点がある．既に，日本語に対しての有効性が確認されているので，本論文では本手法の多言語への汎用性を確かめるために，中国語に対する性能評価実験を行った．実験の結果，90%以上の平均正分割率が得られ，本手法が中国語に適用できることが確認された．このことから，本手法が多言語へ適用できる可能性が示された．

キーワード　中国語，多言語，帰納的学習，単語分割

1. まえがき

分かち書きをしない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，及び音声認識などはすべ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その有効性を確認している [1]．本論文は，多言語に対しての本手法の汎用性を確かめるために本手法を中国語に適用し，その有効性を実験により確認したものである．

中国語の表記法は漢字を羅列するだけで，単語と単語との切れ目を示さない．また，漢字の大部分が表意文字であり，一つの漢字が一つの音節をもち，一つの概念をもつことが多い．単語は一つあるいは複数の漢字で構成される．中国語は語形変化のない言語であり，単語の品詞が使い方によって異なることも多い．例えば動詞は副詞としても，名詞としても用いられることが多い．中国語には日本語のような主語，述語などを明らかにする助詞に相当するものがないので，単語を一定の規則に従って並べることによって正しい文が生成される．複合語の場合は，複合語に含まれる各単語は基礎的文法に従って成立し，修飾と被修飾部分の組合せによって単語となる．文脈によって単語の切れ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．更に，“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，更に一つの単語として存在し得る．このような中国語の特徴によって，単語分割では複数の分割可能性をもつあいまい分割が生じやすい．更に，科学技術の進歩による言語の変化と専門用語，技術用語などの出現によってすべての単語を辞書に収録することも困難である．したがって，中国語の単語分割ではあいまい分割と未知語の処理が主な問題となる．

† 北海道大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkaido University, Sapporo-shi, 060–8628 Japan

中国語の単語分割に関する研究はこれまでいくつか行われてきたが，大別して規則に基づく手法 [3]〜[5]，統計的な手法 [6]，規則辞書と統計情報を融合する手法 [7], [8] がある．文献 [4] ではあいまい分割が生じる種類と原因を分析し，分割規則をまとめる方法を用いてあいまい分割を処理している．規則に基づく手法はあらかじめ単語辞書，あいまい分割を処理する規則を用意する．規則に基づく手法の正解率は辞書の大きさと規則の量に依存する．また，規則の抽出，整理及び更新には膨大な労力がかかる．文献 [6] では統計的な手法を用いて辞書，タグ付きコーパスを使わず漢字間の隣接情報で単語の境界を決定しているが，2 文字だけの単語のみを対象としている．文献 [7] で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手法が提案されている．文献 [8] では規則辞書と統計情報を融合する手法を用いて新聞の文書中の人名を認識する方法が提案されている．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．しかし，一般にスパースネスの問題を避けて高精度な統計モデルを使用するためには大規模なデータが必要である．また，規則辞書と統計的な情報を融合する手法では規則辞書に基づいて統計的な情報で規則を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．また，規則に基づく手法，規則辞書と統計的な情報を融合する手法はあらかじめ単語辞書，タグ付き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，大規模なタグ付き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，タグ付き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行う単語分割方法も提案されている [9]．この手法は隣接漢字の相互情報量で単語の境界を決定している．文献 [10] では n-gram を計算することにより日本語文中の漢字列を単語に分割する手法を提案している．しかし，中国語に対しての評価実験は，実際には行われていない．また，文献 [11] では形態素解析処理の言語に依存した部分を考察し，その部分をできる限り共通化して内部処理のコンポーネント化により複数の言語で共通に利用できる手法を提案している．文献 [12] では単語分割ルールを用いて分割候補のリストを作成し，文字間接続確率最小法を用いて正しい分割候補を選択する手法を提案している．しかし，この手法の分割ルールは日本語の字種変化の情報を利用するために日本語に依存する．

我々の提案する手法は，あらかじめ単語辞書，規則辞書，タグ付きコーパスなどを準備することは必ずしも必要がなく逐次的に正解を提示するオンライン学習の一つである．なお，本論文中では単語分割実験の正解の基準として Sinica corpus の単語分割情報を用いた．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．本手法は字面情報を利用し，文書に頻繁に出現する文字列を再帰的に 2 段階で共通部分，差異部分を抽出する．異なる抽出状況により抽出された共通部分，差異部分を単語としての確実性の高い順に獲得する．本論文では文字列より抽出されたすべての共通部分，及び差異部分を WS（Word segment）と呼ぶ．WS は抽出状況により S1，S2，S3 に分類して辞書に登録する（**3.1**）．そして，WS を用いて文を単語に分割する．ユーザが分割結果の誤りを校正する．この校正過程で単語として確認された WS は CW（Correct word）と呼ぶ．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 CW，WS の正分割度数，誤分割度数，出現頻度などの情報を更新する [13]．本論文で正分割数とは分割結果において人手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，人手によって正しく分割したときに用いられた単語数である．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．正分割度数，誤分割度数とは分割に用いた辞書項目の正分割回数，誤分割回数である．共通部分と差異部分を再帰的に抽出するとは入力文の分割が終わるまで繰り返して共通部分と差異部分を抽出し，抽出された共通部分から更に共通部分及び差異部分を抽出するということである．なお，本論文で帰納的学習とは文書中に重複して出現する文字列を再帰的に抽出する過程である．

以下 **2.** で本手法の概要を述べ，**3.** で本手法の詳細な処理過程について述べ，次に **4.** で本手法の有効性を評価するために行った実験について述べる．続く **5.** で実験の考察結果について述べ，最後に **6.** でまとめと今後の課題について述べる．

2. 概　　要

図 1 に本手法の流れを示す．まず，文を入力して，既に辞書に登録されている CW，WS を用いて文書を単語に分割する．ここで分割されなかった部分文字列

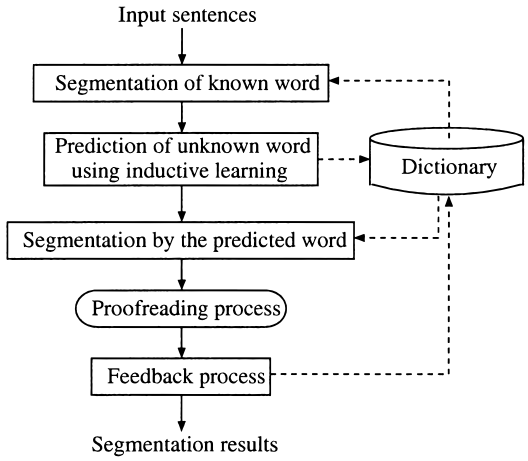

図 1 処理過程
Fig. 1 Process.

について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再帰的に抽出することにより行われる．抽出過程は文書から重複して出現する文字列を再帰的に抽出する過程と抽出された文字列から更に共通部分，差異部分を抽出する過程の 2 段階で行われる．抽出された文字列すなわち WS を単語としての確実性の高い順に分類する．抽出された WS は抽出状況によって単語としての優先順位が決定され，辞書に登録される．そして，推測された WS で分割を行う．ユーザは分割結果の正誤を判断し，分割結果の誤りを校正する．システムは分割結果と校正により得られた正しい結果を比較することにより，辞書に登録されている CW と WS の正分割度数，誤分割度数及び頻度の情報を更新する．更に，校正で得られた正しい分割結果を用いて正しい単語を CW として辞書に登録する．

このような処理を繰り返すことによって，システムは更新された辞書を用いて新しい入力文を分割し，正分割率が徐々に上昇していく．

3. 処理過程

3.1 獲得された語による分割

それまでに辞書に獲得された CW，WS を用いて文を単語に分割する．CW は校正処理により単語として確認されたものである．既に獲得された CW と WS で分割されなかった部分文字列に対して新しく WS を獲得することにより未知語を推測し分割を行う．WS は抽出された共通部分 S1，再抽出された共通部分 S2，残りの差異部分 S3 で構成され，単語としての確実性の高い順に S2，S1，S3 である（**3.2**）．そして，S2，S1，S3 の優先順位で用いて分割を行う．処理手順を以下に示す．

（1） 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すべての分割候補のリストを作成する．分割は文の先頭から行う．ただし，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の値の大きい順に決定し，分割を行う．

（2） 同じゆう度評価関数の値の分割候補が複数個存在する場合，誤分割度数の小さいものを選択する．更に，同じ誤分割度数の分割候補が複数個存在する場合，正分割度数の高いものを選択し，以下，順に出現頻度の高いもの，文字数の多いもの，分割位置が一番前のものの順に決定し，分割を行う．

（3） 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，S2，S1，S3 の順に用いて分割を行う．また，分割に重複が存在する場合，ゆう度評価関数の値，及び（2）の手順に従って正しい分割候補を決定して分割を行う．

ここでゆう度評価関数 LEF（Likelihood Evaluation Function）を以下に示す．

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

FR（Frequency of appearance），CS（Frequency of correct segmentation），ES（Frequency of erroneous segmentation），LE（Length of CW or WS）は登録された CW あるいは WS の正分割度数，誤分割度数，出現頻度，単語のバイト数である．α，β，γ は重み係数である．このゆう度評価関数は出現頻度が高く，正分割度数が高く，誤分割度数が低く，更に文字数の多い CW，WS を優先的に用いることを表している．また，ゆう度評価関数はこれらの特徴量の線形和を採用したが仮に用いたものであり検討の余地があると考えられる．

3.2 未知語の推測

獲得された CW，WS で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 2 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

中國民營科技 企業經過多年的 發展，目前正歩入高速 發展 的時期．民營科技 活動，已經覆蓋了 國民 經済主要行業，成為中國發展 高 科技 産業的生力軍．

図 2　中国語文書の例
Fig. 2　An example of Chinese text.

表 1　辞書の構造
Table 1　Construction of the dictionary.

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

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．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．図 2 に中国語の文書の例を示す．この図 2 により次に未知語の推測方法を述べる．

3.2.1　共通部分の抽出

この段階で抽出される文字列を S1（Segment one）と呼ぶ．抽出条件を以下に示す．

（1） 2 文字以上の文字列が文書中に重複して出現する場合，S1 として抽出する．図 2 の文書において，抽出された S1 としての単語候補は “中國”，“民營科技”，“發展”，及び “國民” である．

（2） 抽出された S1 に含まれ，かつ他の S1 のどこにも含まれていない出現位置が少なくとも一つある文字列も S1 として抽出される．例えば，“科技” は抽出され，S1 とする．

3.2.2　高次共通部分の抽出

抽出された共通部分に他の共通部分が含まれる場合には，更に共通部分と差異部分を抽出する．これは，一つの共通部分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．この S1 からの共通部分の再抽出は高次共通部分の抽出という．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．高次共通部分の抽出条件を以下に示す．

（1） 共通部分 S2 の抽出と残りの差異部分 S3 の処理：抽出された S1 同士の共通部分が存在する，あるいは，S1 に別の S1 が含まれていて，更に別の S1 に含まれない出現位置が少なくとも一つ存在する場合，この S1 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．例えば，“科技（S1）” は “民營科技（S1）” に含まれ，“民營科技” から “科技” を抽出できる．よって，“科技” を抽出して S2 に所属させ，残りの差異部分 “民營” は S3 に所属させる．

（2） 1 文字単語の処理：1 文字が分割済みの単語で囲まれている場合，この 1 文字を単語として抽出する．抽出された 1 文字の単語は S2 に所属させる．図 2 において，“高” は “發展高科技” において “發展” と “科技” で囲まれており，両側は分割済みのため，“高” を抽出して S2 に所属させる．

抽出した共通部分，差異部分を WS といい，単語として確実性の高い順に辞書に登録する．このようにして抽出された WS は，単語としての確実性の高い順に S2，S1，S3 である．

3.3　辞書の構造

帰納的学習で抽出された WS を S1，S2，S3 に分類して，辞書に登録する．登録する際に同時に WS の正分割度数，誤分割度数，頻度，単語のバイト数及びその分類も登録する．表 1 に辞書の構造を示す．ここで，CL は登録された WS の分類（Classification）である．CW に所属しているものは校正処理で確認された正しい単語である．また，辞書は CW，及び S1，S2，S3 で構成される．

3.4　推測された語による分割

既に登録された CW と WS を用いて分割されなかった部分に対して，新しく推測された WS を用いて単語として確からしさの高い順に分割が行われる．ここで推測された WS を S2，S1，S3 の順で用いて分割されなかった文字列とマッチングして分割候補のリストを作成する．そして，単語として確実性の高いものを優先的に用いて分割を行う．具体的な処理過程を以下に示す．

- S2，S1，S3 の順で WS を用いて文とマッチングして分割候補のリストを作成し，分割を行う．
- 分割候補が複数個存在する場合，ゆう度評価関数の値が大きいものを優先的に用いて分割する．
- ゆう度評価関数の値が同じ分割候補が複数個ある

場合の処理は **3.1** の獲得された語による分割の（2）と同様である．

3.5 フィードバック処理

ユーザが分割結果の正誤を判断し，分割結果に誤りが含まれている場合，ユーザが誤りの箇所を校正する．次に，システムは分割結果と校正した正しい分割結果を比較することにより，未分割部分の単語を辞書に自動的に登録し，辞書に登録されている単語候補の正分割度数，誤分割度数，頻度及び分類を更新する．正分割の場合，分割に用いた CW あるいは WS の正分割度数を増加させる．誤った分割の場合，分割に用いた CW あるいは WS の誤分割度数を増加させる．CW，WS のゆう度を更新する方法は以下のとおりである．

（1） 正分割結果のゆう度の更新

正分割の場合，正分割に用いた CW，WS の頻度と正分割度数を増加させる．更に，正分割に用いた WS の分類を CW とする．

（2） 誤分割結果のゆう度の更新

誤分割に用いた CW，WS の誤分割度数を 1 増加させる．更に以下の処理を行う．

（a） 正しい単語が辞書にない場合，正しい単語の頻度 FR を 1 にして，CW として辞書に登録する．

（b） 正しい単語あるいは WS が辞書にある場合，システムはその単語あるいは WS の頻度 FR を 1 増加する．更に WS の所属を CW とする．

（3） 未分割結果の処理

既に登録されている CW，WS を用いて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 で分割する．このようにしても分割されなかった部分文字列に対してフィードバック処理で校正済み結果と分割結果とを比較することにより正しい単語を辞書に自動的に登録する．登録された単語の頻度を 1 にして，分類を CW にする．

図 3，図 4 に単語分割結果と校正された分割結果の例を示す．図 3 と図 4 中の中括弧で囲んでいる部分は分割され，かつ獲得された単語を意味する．この場合，正分割数は図 3 と図 4 で同じ部分の数である．未分割数は未分割文字列を，人手によって正しく分割したときに用いられた単語の数（{ 經過 }），1 である．したがって誤分割数は人手によって正しく分割するときの総単語数から正分割数と未分割数を除いた数，2 である．図 3 において { 中 }，{ 國民 }，{ 營 } と “經過” 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．次にフィードバック処理において登録された CW，WS のゆう度の更新過程を述べる．

{ 中 }{ 國民 }{ 營 }{ 科技 }{ 企業 }經過{ 多 }{ 年 }{ 的 }{ 發展 }，{ 目前 }{ 正 }{ 歩入 }{ 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
Fig. 3 Result with errors.

{ 中國 }{ 民營 }{ 科技 }{ 企業 }{ 經過 }{ 多 }{ 年 }{ 的 }{ 發展 }，{ 目前 }{ 正 }{ 歩入 }{ 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 校正された単語分割結果
Fig. 4 Corrected result.

- 正分割結果の処理
  “科技”，“企業” などの正しい分割に用いた CW，WS の頻度を 1 増やす．CW である場合，正分割度数を 1 増加する．WS である場合，CW に所属させる．
- 誤分割結果の処理
  “中”，“國民”，“營” などの誤分割に用いた CW，WS の誤分割度数を 1 増加する．更に，正しい単語が辞書にあれば，頻度を 1 増やし，WS である場合，CW に所属させる．正しい単語が辞書になければ，頻度 1 にして，CW として辞書に登録する．
- 未分割結果の処理
  “經過” は未分割部分である．“經過” の頻度を 1 にして，CW として登録する．

4. 評価実験

本手法の多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，中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) に示

表 2　予備実験結果
Table 2　Results of preliminary experiment.

係数	α	1	1	1	1	1	**1**	1	1	0	5	10
	β	60	70	80	70	70	**70**	70	70	70	70	70
	γ	1	1	1	30	40	**50**	60	70	50	50	50
結果	正分割率 [%]	77.91	77.93	77.92	78.11	78.14	**78.15**	78.13	78.11	78.03	78.12	78.08
	誤分割率 [%]	20.92	20.90	20.91	20.72	20.69	**20.68**	20.70	20.72	20.80	20.71	20.75
	未分割率 [%]	1.17	1.17	1.17	1.17	1.17	**1.17**	1.17	1.17	1.17	1.17	1.17

す正分割率，誤分割率，未分割率を用いた．

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

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

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

4.1 予備実験

まず，ゆう度評価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．250 文の中国語の経済学の文書（約 7,000 単語）を用いて，欲張り法（greedy method）[14] でゆう度評価関数の係数を変化させ最適な正分割率が得られる値を求めた．まず α，β に初期値を与え，γ を変化させ，正分割率が最大となる γ の値を求める．次に β の値を変化させ正分割率が最大となる β の値を求める．そして，α を変化させ正分割率が最大となる α の値を求める．このような操作を繰り返して正分割率が最大となる係数の値を最適な係数の値とする．実験結果の一部を表 2 に示す．この実験結果から正分割率が最大となった最適な係数は $\alpha = 1$，$\beta = 70$，及び $\gamma = 50$ となった．

4.2 実験データ

実験データとして Sinica Corpus(注1)から建築学 145,727 単語，経済学 113,000 単語と電子工程学 116,110 単語の文書のデータを用いた．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．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．電子工程学には電子工学，通信工程及び器械工程，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実験を行った．

4.3 実験手順

実験は辞書が空の状態からはじめ三つの分野の文書を 1 分野ずつ，一つの分野の文書に対して約 100 単語の文書を 1 段落ずつ入力して実験を行う．システムは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，辞書に登録する．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分割結果とを比較することにより，辞書に登録された CW，WS の情報を更新する．

なお，本手法が言語に依存しないという特徴，及びユーザに適応できることを示すために実験の初期状態は辞書を空に設定した．

4.4 実験結果

実験結果を表 3 と図 5 に示す．図 5 は正分割率，誤分割率，未分割率の推移を表している．表 3 は 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における平均分割結果と全体の平均分割結果を示している．

5. 考　　察

5.1 有効性

表 3 に示されるように 3 分野それぞれの平均正分割率は 89.3%，91.2%，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い，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 95%以上になったが，その後辞書に登録された CW，WS が増加するのに伴い，あいまい分割が原因で誤分割率が大きくなっている．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．

また，局所的に変化する箇所が見られ，その部分の

(注1)：Sinica Corpus,
http://www.sinica.edu.tw/ftms-bin/kiwi.sh

表 3 実験結果
Table 3 Results of experiment.

	architecture	economics	engineering	average
Number of words	145,727	113,000	116,110	374,837
正分割率 [%]	89.3	91.2	91.9	90.6
誤分割率 [%]	10.1	8.4	8.1	9.0
未分割率 [%]	0.6	0.4	0.0	0.4

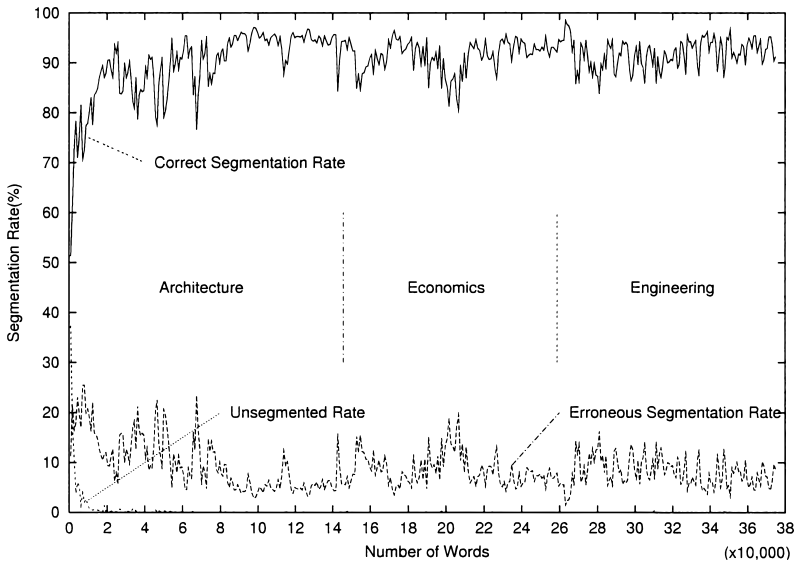

図 5 正分割率，誤分割率，未分割率の推移
Fig. 5 Changes in correct segmentation rate, erroneous segmentation rate and unsegmented segmentation rate.

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 5 に示される誤分割率の推移は，処理される単語数の増加に従い誤分割率が下がっていることを示している．更に，図 5 中の未分割率の推移は未知語を推測する能力を表している．実験のはじめは辞書が空なので，分割は推測された WS により行われたため，未分割率は高かった．しかし，辞書が自動的に生成されるに従って未分割率が大幅に減少し，約 20,000 単語を処理した後，総平均未分割率は 0.4%であった．表 4 と図 6 に処理された単語数の増加に伴う，結果の誤りを校正する回数の変化を示す．辞書が空の場合の回数と 111,000 単語を処理した後の回数とを比較すると，98 回から 44 回に減少している．

また，本手法の学習機能について比較実験を行った．実験 1 は本論文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，分割の誤りを校正して分割を行った実験である．実験 2 は共通部分と差異部分を抽出せず，文を入力して，分割の誤りを校正して分割を行った実験である．実験には建築学分野のデータと経済学分野のデータ，総単語数 180,000 単語を用いた．実験 1 の場合の正分割率と実験 2 の場合の正分割率の推移，及び実験 1 の実験 2 に対する正分割率の改善率の推移をそれぞれ図 7，図 8 に示す．ここで改善率を式 (5) に示す．正分割率 1 と正分割率 2 はそれぞれ実験 1 と実験 2 の正分割率である．

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わかる．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わかる．また，図 8 からわかるように実験の最初と分野が変化するときの正分割率の改善率が大幅に増加している．これにより本手法の学習機能の有効性が示された．実験の最初の 20,000 単語を処理したとき

表 4　校正回数
Table 4　Proofreading times.

Range of Word	0–1,000	10,000–11,000	20,000–21,000	30,000–31,000
Corrected Times	98	132	84	96
Range of Word	40,000–41,000	50,000–51,000	60,000–61,000	70,000–71,000
Corrected Times	89	68	44	45
Range of Word	80,000–81,000	90,000–91,000	100,000–101,000	110,000–111,000
Corrected Times	87	45	36	44

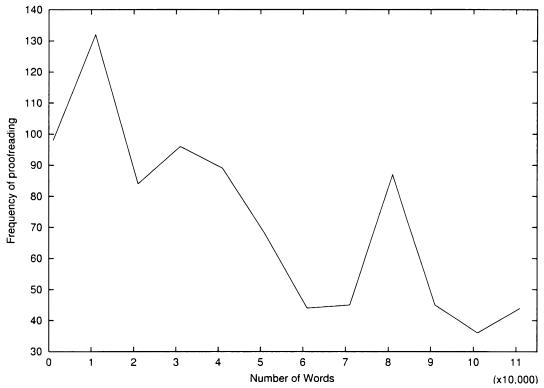

図 6　校正回数の推移
Fig. 6　Change in times of proofreading.

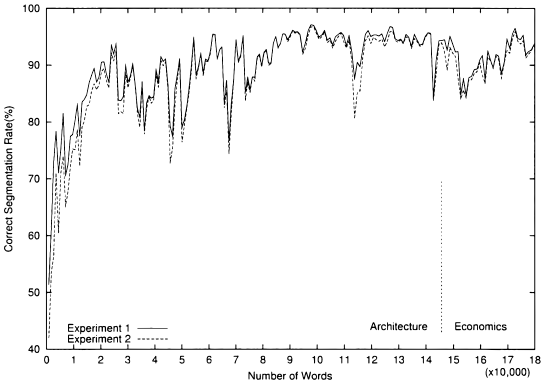

図 7　実験 1 と実験 2 の正分割率の推移
Fig. 7　Change in correct segmentation rate of experiment 1 and experiment 2.

の正分割率の平均改善率は 20.4%である．実験の最初は辞書が空のためシステムが未知語を推測することにより分割を行う．また，実験の最初及び分野の変化のほかに改善率が局所的に変化する箇所が見られ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．

分野及び分野内での文書が変化するとき未知語が出現するために，実験 2 の正分割率が下がっている．これに比べて実験 1 の方が急速に異なる文書，分野に適応できることが図 7 により確認された．よって本システムのもつ未知語を処理する能力の有効性が確認された．

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

5.2　誤りの原因

文書中のはじめの 1,000 単語の段落と 360,000 単語から 361,000 単語の間の段落の文書を取り出して考察を行った．表 5 に誤分割の原因を示す．実験のはじめは辞書が空なので分割候補が足りないため，既登録の CW，WS を用いて誤分割が行われている．例えば，出口（export）が既に登録されているが，出口到

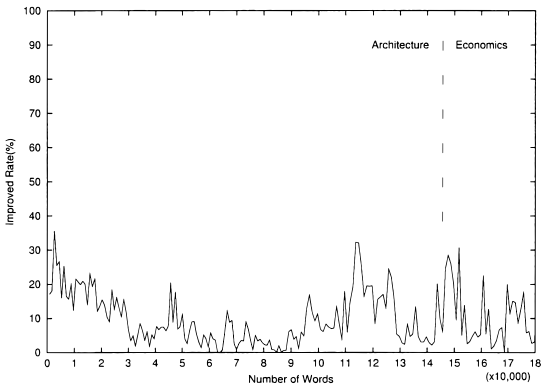

図 8　改善率の推移
Fig. 8　Change in improved rate.

表 5　誤分割原因の割合
Table 5　Ratio of error segmentation.

		0–1,000 words	360,000–361,000 words
割合	未登録語による誤り	15.0%	0.8 %
	あいまい分割による誤り	1.5%	6.0%
	コーパスの揺れによる誤り	0.5%	0.4%
合計	平均誤分割率	17.0%	7.2%

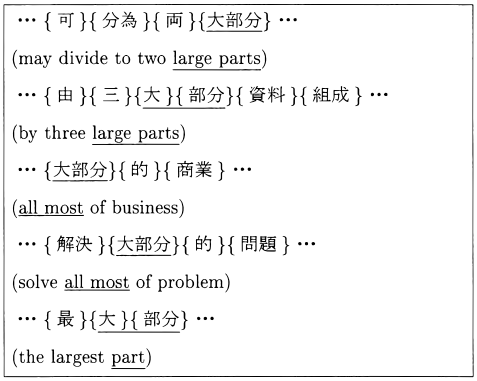

図 9　あいまい分割の例
Fig. 9　Example of ambiguity segmentation.

（export to），出口商（export businessman）と出口額（volume of export trade）などは登録されていないため，分割結果が { 出口 } 到 …，{ 出口 } 商 … と { 出口 } 額 … になった．これらの正しい分割結果は { 出口到 }，{ 出口商 } と { 出口額 } である．この原因で誤分割率は表 5 に示したように実験のはじめ 1,000 単語で平均 17%，終わりに近い 1,000 単語で平均 7.2%であった．同様に，それぞれの誤分割中の未登録語による誤り，あいまい分割による誤り及びコーパスの揺れによる誤りの割合は未登録語による誤りが 15%と 0.8%，あいまい分割によるものは 1.5%と 6.0%，コーパスの揺れによるものは 0.5%と 0.4%であった．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あいまい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．あいまい分割の例を図 9 に示す．図 9 で文字列 “大部分” は文脈によって分割結果が異なる．このようなあいまい分割に対して本手法によりほとんどの部分を正しく分割することができたが，更に分割精度を向上させるためには共起情報や文脈情報などを利用することが考えられる．

5.3　他手法との比較

本手法と文献 [9] との比較実験を行った．文献 [9] のシステムを実際に作成して本論文の実験データを用いてその実験結果と比較検討を行った．トレーニングデータとして本論文で用いた建築学の分野の文書（145,727 単語，512 KB）を用いた．実験はトレーニングデータの 130,000 単語から 140,000 単語の間の連続的な 100 文を取り出して実験を行った．正解の基準として Sinica corpus のタグ付きデータを使用した．表 6 に 100 文の比較実験結果を示す．なお，本手法の正解数は正しく分割された単語数であるのに対して，文献 [9] の正解数は正しく分割された位置の数である．文献 [9] の評価方法で評価した本手法の正分割率と本手法の評価方法で評価した本手法の正分割率を表 6 に示す．表 6 より本手法の単語分割結果の正分割率は文献 [9] より高い．しかし，本手法はオンライン学習の一種であるのに対して文献 [9] は統計的な手法である．方法，実験条件などが異なり単純に比較することはできないが本手法の性能を評価するための参考として比較実験を行った．

表 6　共通データの比較実験結果
Table 6　Results of comparison experiment.

実験結果	文献 [9] の結果	本手法の結果	本手法の結果
評価方法	文献 [9]	文献 [9]	本論文
正分割率 [%]	89.6	97.8	94.6

6. むすび

本論文では字面情報から帰納的学習を用いて未知単語を推測することにより文を単語に分割する手法を提案した．そして，本手法を中国語に適用し，多言語へ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．三つの分野の文書を用いて 90.6%の平均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．更に，分野の変化に追従し，どの分野にも適応できることが確認された．また，本手法は既に日本語を処理できることが確認されている上に，本論文によって中国語を処理できることも確認された．すなわち本手法が言語に依存せず多言語に対応できるという可能性が確認された．

今後は分割あいまい性の解消により分割精度を更に向上させること及び，帰納的学習を用いて品詞を付けることを計画している．

文　献

[1] 荒木健治，栃内香次，“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，” 信学論（D-II），vol.J75-D-II, no.7, pp.1213–1221, July 1992.
[2] 香坂順一，中国語の単語の話，光生館，東京，1971.
[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, April 1996.
[4] 顧　萍等，“漢語自動分詞的近隣匹配算法及其在 QHFY 漢英機器翻訳系統中的実現，” 計算語言学研究与応用，pp.132–138, Nov. 1993.
[5] K.J. Chen and S.H. Lin, “Word identification for mandarin Chinese sentences,” Proc. Coling 92,

pp.101–107, Nantes, France, Aug. 1992.
[6] R. Sproat and C. Shih, “A statistical method for finding word boundaries in Chinese text,” Computer Processing of Chinese and Oriental Languages, vol.4, no.4, pp.336–351, 1990.
[7] 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.
[8] 宋　柔等，“基于語料庫和規則庫的人名識別法,” 計算語言学研究与応用，pp.150–154, Nov. 1993.
[9] M. Sun, D. Shen, and B.K. Tsou, “Chinese word segmentation without using lexicon and hand-crafted training data,” Proc. Coling-ACL’98, pp.1265–1271, Montreal, Quebec, Canada, Aug. 1998.
[10] R.K. Ando and L. Lee, “Mostly-unsupervised statistical segmentation of Japanese,” NAACL’2000, pp.241–248, 2000.
[11] 山下達雄，松本裕治，“言語に依存しない形態素解析の枠組,” 自然言語処理，vol.7, no.3, pp.39–56, July 2000.
[12] 飯塚泰樹，“接続確率最小法による教師なし単語分割,” 情処学自然言語処理研報，2000-NL-139, pp.33–40, 2000.
[13] Z. Wang, K. Araki, and K. Tochinai, “Word segmentation method using inductive learning for Chinese text,” Proc. IASTED International Conference, Artificial Intelligence and Soft Computing, pp.452–458, July 2000.
[14] 浅野孝夫，今井　浩，計算機とアルゴリズム，オーム社，東京，2000.

（平成 12 年 10 月 23 日受付，13 年 5 月 14 日再受付）

王　　忠建　（学生員）

平 6 中国遼寧工程技術大学大学院修士課程了．現在，北大大学院博士課程在学中．自然言語処理の研究に従事．情報処理学会会員．

荒木　健治　（正員）

昭 57 北大・工・電子卒．昭 63 同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手．平 1 同講師．平 3 同助教授．平 10 同教授．現在，北大大学院工学研究科電子情報工学専攻助教授．自然言語の機械学習と機械翻訳に関する研究に従事．言語処理学会，日本認知科学会，人工知能学会，情報処理学会，ACL，AAAI，IEEE 各会員．

栃内　香次　（正員）

昭 37 北大・工・電気卒．昭 39 同大大学院修士課程了．現在，同大大学院工学研究科電子情報工学専攻教授．主として音声情報処理，自然言語処理の研究に従事．工博．情報処理学会，日本音響学会各会員．