自然言語処理 129－12
（1999. 1. 21）

# 表層解析に基づく点字用日本語分かち書きへの事例ベースの適用

高木 喜次† 小野 智司† 宮下 和雄‡ 西原 清一†
†筑波大学 電子・情報工学系
‡電子技術総合研究所

日本語文書を点字に翻訳する問題を取り上げ、過去の分かち書きの誤りを蓄積した事例ベースを、ルールベースに補完的に適用する枠組みを提案する。点字翻訳における分かち書き問題は、全ての規則を前もって記述することが難しく、さらに多くの例外が存在するために、従来のルールベースとユーザ辞書を組み合わせて処理する手法では、それらすべてを規則として記述することは難しい。本手法では、知識適用の結果、誤って分かち書きした箇所を事例として自動獲得する。それ以後の分かち書きにおいては、ルールベースによる分かち書きを行った後、入力文字列と獲得した事例との類似度を計算し、事例に基づいて分かち書きを修正する。本手法を情報処理関連のテキストに適用して、その有効性を確認し、さらにその結果に対する解析を行った。

# A Japanese Sentence Segmentation Method for Braille Based on Surface Analysis and Case-Based Reasoning

Yoshitsugu Takagi† Satoshi Ono† Kazuo Miyashita‡ Seiichi Nishihara†
† Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1, Tennohdai, Tsukuba, Ibaraki 305-8573, Japan
‡ Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba, Ibaraki 305-8568, Japan

We propose a Japanese sentence segmentation method based on the surface analysis, which makes use of case-based knowledge as well as rule-based one to translate Japanese text into Braille with high reliability. Japanese sentence segmentation is not an easy task when only rule-based knowledge and a user dictionary are available because of many exceptions to each rule. In our method, those exceptions are stored as the case-base, in which each element, or a case, is a fragment of experiential knowledge. Our system is expected to give higher quality of segmentation results as the case-base grows. After giving an algorithm to measure the similarity between a case and a given character string, we perform some experiments to evaluate the method.

## 1. はじめに

近年、バリアフリーというキーワードのもと、各種福祉機器が開発され、それらを用いて視覚障害者が社会で活躍する機会が増えてきている。しかし、視覚障害者にとっての情報源は、音声情報や点字化された文書に限られるため、その情報量は晴眼者に比べて決定的に不足している。このため、既存の電子化されたテキストなどの点字化が必要とされており、様々な自動点字翻訳システムが公表されている。しかし、点字翻訳（以後、点訳とする）をコンピュータによって自動的に行うには様々な問題があるため、既存の手法を用いても満足できる結果を得ることはできず、点訳ボランティアは、自動点訳システムを用いずに点字を入力しているのが現状である。

点訳において、最も難しいとされている分かち書き問題には、文法的な規則のほかに意味や拍数を考慮する必要がある曖昧な規則が存在する。これによって、一般の形態素解析などによる情報を用いて規則を記述することが難しい。また、点訳の対象となる文章の分野ごとに異なる規則を使い分ける必要があるため、そのすべての規則を列挙することは困難である。さらに、例外が数多く存在するという自然言語の特質上、分かち書きのすべての例外的な規則を、事前に列挙するのは不可能であり、全規則を事前に導出・網羅した規則集を作成することは難しい。その対策として、記述しきれない規則を、経験の積み重ねによって随時獲得していく学習の枠組みが必要であると考えられる。

本研究では、ルールベースに対し事例ベースを補完的に併用する枠組みを提案する。

本手法は、まず知識適用による自動分かち書きの誤りを事例として獲得する。次に表層解析によって得られる情報に基づいて類似度計算を行ない、類似度が閾値を超えたものについては事例を適用し、分かち書きの精度を向上させるものである。

以下2章では、点訳の概要と分かち書きにおける問題点、既存の手法について述べる。3章では、提案する手法についてその特長や処理の流れ、事例の表現方法や類似度の計算方法を詳しく説明する。4章では、情報処理関連のテキストを用いて行った実験についての解析結果について述べ、さらに5章で今後の課題について述べる。

## 2. 点字翻訳の概要と問題点

### 2. 1 点字翻訳について

日本語の点訳においては、漢字をかなに変換するとともに、単語と単語の間に空白を挿入する分かち書きを行う必要がある。漢字かな混じり文の例文を点訳する様子を図1に示す。

欧米諸言語のように分かち書きされ、単語と単語の間に空白が挿入されている言語とは異なり、日本語は字種が豊富である上に単語と単語が分かれていない。加えて、一般の漢字かな混じり文は漢字とかなの使い分けで文を構成しているのに対し、点字は表音文字体系である。このため、一定のルールに従って単語と単語の間を区切らないと読みにくいばかりでなく意味も正確に伝わらない。区切り方によって意味が異なる例を図2に示す。

したがって点訳する際には、漢字をかなに変換するだけではなく、分かち書きも行う必要が

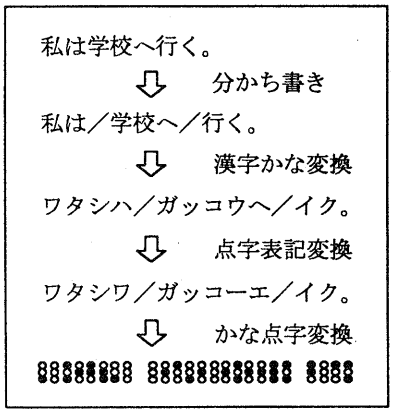


図1： 点字翻訳の手順

例）キョーカヨーカモシレナイ。

キョー／カヨーカモ／シレナイ。
⇨ 今日火曜かも知れない。

キョーカヨーカモ／シレナイ。
⇨ 強化用かも知れない。

図２：分かち書きによって意味が変わる例

ある。日本語の点訳用分かち書きにはさまざまな例外が存在する。それら全てをルールで表現し、既存のルールとの強弱関係を考慮し、優先点数を決定する作業は非常に効率が悪く、ルールの追加による精度の向上は難しいと考えられる。このため本研究では、知識を自動的に獲得して分かち書きの精度を向上させる、事例ベースを用いた枠組みを提案する。

## ２．２　分かち書き処理における問題点

点字翻訳における分かち書きの規則の例を表１に示す([1]より抜粋)。

表１：分かち書きの規則の例

| | 分かち書き規則 |
|---|---|
| 規則 1-1 | 接頭接尾語と結びついている複合語は区切らずに書く |
| 規則 1-2 | 接頭接尾語であっても意味の理解を助ける場合には発音上の切れ目を考慮して区切って書く |
| 規則 2-1 | 長い複合語で内部に３拍以上の自立可能な意味の成分が二つ以上ある場合は、原則としてその境目で区切って書く |
| 規則 2-2 | 自立可能な語が連接した複合語であっても、分ければ独自の意味が失われる複合語は区切らずに書く |
| 規則 2-3 | 自立可能な２語が連接した複合語であっても、後半の語頭に連濁を生じる場合は区切らずに書く |

表１の規則は学校文法などの一般的な文法情報に基づくものであるが、意味や発音を考慮しなければならない曖昧な規則もあり、一般的な形態素解析に基づく情報を用いても、このすべての規則を表現するのは困難である。

## ２．３　ルールベースに基づく分かち書き処理の概要

まず、筆者らは分かち書きの処理を、ルールベースによる自動分割と対話処理の２段階で行う手法 [3][4][7][8] を構築した。その処理の手順を図３に示す。

その手順としては、まず入力された漢字仮名混じり文に対し表層解析を行ない、その情報に基づいて、各文字間に対し if-then 形式で記述されたルールを適用する。注目する箇所でルールが競合する場合は、各ルールに設定した優先点数に基づいて適用するルールを決定する。ルールの例を表２に示す。

ここでは、辞書が大規模である形態素解析を行わず、表層解析を採用することで、実装の手間の軽減と分かち書き処理時間の短縮を図った。表層解析とは、字面や字種などの表層的な情報のみの解析を行うものであり、本研究では、これに加え７種類の小規模な（合計約 0.5MB）テーブルを用いて助詞、漢字熟語、ひらがな書き自立語などの判定も行っている。

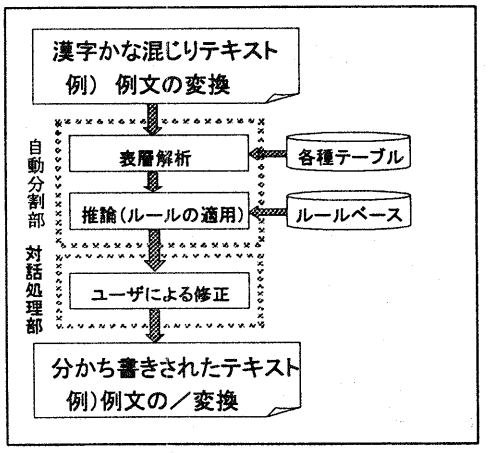


図３：従来手法の処理の流れ（ルールベースのみ）

表２：　ルールの例

```
ルール(1－1,9,
    [if,   [後ろの文字の字種は,[句読点]]]
    [then, [区切り方は[区切らない]]])
ルール(1－2,10,
    [if,   [前の文字は,[句点]]]
    [then, [区切り方は[2回区切る]]])
ルール(1－3,10,
    [if,   [前の文字は,[読点]]]
    [then, [区切り方は[1回区切る]]])
```

```
ルール(5－1,7,
    [if,   [後ろの文字は,[ひらがな書き自立語の1文字目]]
           [後ろの文字は,[前後を区切る種類のひらがな書きの自立語の一部],
           [前を区切る種類のひらがな書きの自立語の一部]]]
    [then, [区切り方は[1回区切る]]])
ルール(5－2,5,
    [if,   [前の文字は,[ひらがな書き自立語の最後]]
           [前の文字は,[前後を区切る種類のひらがな書きの自立語の一部],
           [後ろを区切る種類のひらがな書きの自立語の一部]]]
    [then, [区切り方は[1回区切る]]])
```

## ２．４ ルールベースに基づく分かち書き処理の問題点

ルールベースに基づく手法では、分かち書きの精度を向上させる手段として、ルールの追加を行うという方法が考えられる。しかし、日本語の点訳用分かち書きにはさまざまな例外が存在する。それら全てをルールで表現し、既存のルールとの優先点数を考慮し決定する作業は非常に効率が悪く、ルールの追加による精度の向上は難しいと考えられる。このため本研究では、知識を自動的に獲得し、分かち書きの精度を向上させる事例ベースを用いた枠組みを提案する。

# ３．ルールベースと事例ベースに基づく手法

## ３．１　基本方針と特長

今回我々が構築した手法の基本方針と特長を以下に示す。

（ｉ）ルールベースに対して、事例ベースを補完的に併用する。

⇒事前に記述できなかった規則や例外を自動的に導出することができる。

（ii）ユーザの対話から事例を獲得する。

⇒知識適用による分かち書きが誤った箇所をユーザの指摘によって事例として獲得するため、ユーザにとって必要な事例のみを獲得することができる。

（iii）問題解析器として表層解析を用いる。

⇒現段階では、既存のルールベースに基づく手法を基に、知識を自動的に獲得する枠組みを構築することを目的としているため、表層解析を用いる。

## ３．２　提案する手法の処理の流れ

提案する手法の処理の流れを図４に示す。

処理の手順としては、まず入力された漢字かな混じり文に対して、各種テーブルを用いて表層解析を行う。そして、付加された情報に基づいてルールベースを適用し、さらにその後、事例ベースを適用する。その結果が誤っており、ユーザに修正された場合には、その箇所を事例として事例ベースに獲得する。最後に、分かち書きされた漢字かな混じり文を出力する。

ルールベースと事例ベースを併用することにより、既存のルールでは全て記述することができなかった、例外的な分かち書きを事例とし

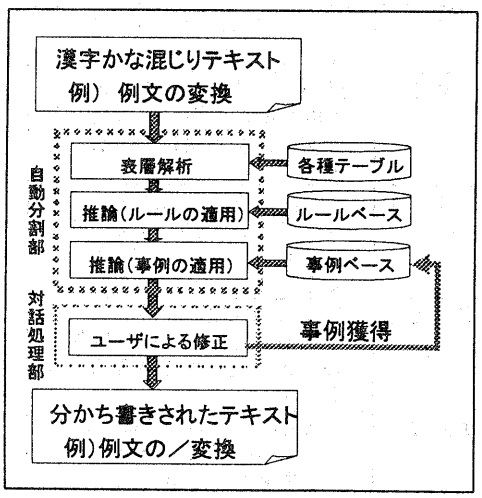


図４：提案する手法の処理の流れ
(事例ベースを併用)

て蓄積し、入力テキストと事例との類似度に応じて、各文字間でのルールベースによる区切りを修正することが可能となる。

### 3．3　表層解析

本手法で用いている表層解析について説明する。ここで表層解析とは、漢字かな混じり文の文面から判別できる字種や、以下に示す表層情報を抽出することである。具体的には、表3に示す7種類のテーブル名が示す情報を表層情報と定義し、文中のどこからどこまでがそれぞれのテーブルの要素と一致したかを調べる。形態素解析では辞書を引いて単語ごとに文法情報を獲得し、文法情報に基づいて連接を詳しく調べる。これに対して、表層解析では、文に出現する文字列の表層情報をテーブルに基づきチェックするだけで、前後の語と語の関係については考慮しない。

本手法の表層解析で用いるテーブルはここに示した7つであり、従来の形態素解析辞書のような文法情報は持っていない。ひらがな書き自立語テーブル以外は全て単語のみからなる。

ひらがな書き自立語とは「しかし」「はっきり」のような、全てひらがなで表記される自立語を示す。これらは表層解析の際、分かち書きの重要な目安となる、助詞との区別がつきにくいうえ、ひらがな書き自立語どうしが接続した際、分かち書きを誤る可能性が高いため、前後における区切り情報もテーブルに保持する。

表3：表層解析に用いるテーブルの一覧

| テーブル名 | 例 | 語数 |
|---|---|---|
| 助詞 | より | 24 |
| ひらがな自立語 | しかし: 前後 | 303 |
| 混ぜ書き語 | 引き算 | 10,826 |
| 漢字2字熟語 | 情報 | 62,512 |
| 漢字3字熟語 | 亜熱帯 | 24,262 |
| 接頭語 | 副 | 54 |
| 接尾語 | 員 | 67 |

分かち書きを「証券業／務を」と誤った際に獲得される事例の例

字面：証券 業務を
字種：漢漢 漢漢ひ
漢字情報：22 210
助詞情報：00 001
：　：

注目する箇所（「券」の直後）における区切り情報：「区切る」

図5：獲得する事例の例

これらのテーブルは、ＥＤＲ日本電子化辞書研究所の「日本語基本単語辞書」から単語を抽出し、ひらがな書き自立語については区切り方の情報を人手により入力した。また、接頭・接尾語については実際のテキストから抽出して作成した。

### 3．4　事例の表現方法と類似度の定義

事例ベース推論において最も重要となるのは事例の表現方法、および事例と推論対象との類似度の定義である。本研究における事例は、属性として表層情報、またクラスとして、注目する箇所で「分かち書きを行うかどうか」という区切り情報で表現される。事例獲得の例を図5に示す。

事例は注目する区切り箇所を中心としてそれぞれ直前、直後の区切りまでを一つの事例として獲得する。そのとき、入力文字列とその表層情報、適用されたルール番号などが事例ベースに格納される。図5においては、“券”の後ろで誤っているため、その前後の正しい区切り位置である“証”の前から“を”までを事例として獲得する。属性値としては「字面」（入力文字列）、「字種」（入力文字列の字種情報）、「その他の表層情報」（表層解析を行う際に、どのテーブルを参照し、どのように一致したかを示す整数値）を獲得する。

図5では、「字面」として“証券業務を”、「字種」として“漢字, 漢字, 漢字, 漢字, ひらがな”

という値を獲得している。また，「その他の表層情報」の例として，漢字情報について説明する。図５の例においては、〈証券〉〈証券業〉〈業務〉の３語が漢字熟語テーブル内の見出しと一致し、その重複回数が属性値として獲得される。

本手法では、入力と事例の類似度は次式により定義し、３種類の類似度すなわち字面に注目した類似度、字種に注目した類似度、その他の表層情報に注目した類似度から計算する。

全体の類似度 ＝（字面の類似度）×
（字種の類似度）×（字面字種以外の類似度）

$$(\text{字面の類似度}) = \begin{cases} 1.0 : \text{完全に一致したとき} \\ 0.9 : \text{完全には一致しないとき} \end{cases}$$

$$(\text{字種の類似度}) = \frac{(\Sigma_{i=1}^{n} \delta_i \times \rho_i)}{\Sigma_{i=1}^{n} \delta_i}$$

$$\rho_i = \begin{cases} 1 : \text{字種が一致したとき} \\ 0 : \text{字種が一致しないとき} \end{cases}$$

$\delta_i$ ： 注目する区切りに近い文字がより重くなるよう設定した、距離の重み

n ： 事例の長さ

$$(\text{その他の表層情報の類似度}) = \frac{(\Sigma_{i=1}^{n} \omega_i \times \sigma_i)}{\Sigma_{i=1}^{n} \omega_i}$$

$$\sigma_i = \begin{cases} 1 : \text{表層情報が一致したとき} \\ 0 : \text{表層情報が一致しないとき} \end{cases}$$

$\omega_i$ ： 注目する文字の字種による重み

ここで字面に注目した類似度とは、入力テキストと比較する事例との文字が完全に一致するかどうか、また字種に注目した類似度とは入力テキストと事例の字種が一致している割合を示している。その他の表層情報に注目した類似度では、入力文字列と事例を表層解析した際の表層情報を比較し、それらが一致する割合を計算する。類似度計算の例を図６に示す。

まず、入力文字列の“注目する区切り”で適用されたルールの番号（ここでは「２７」）を

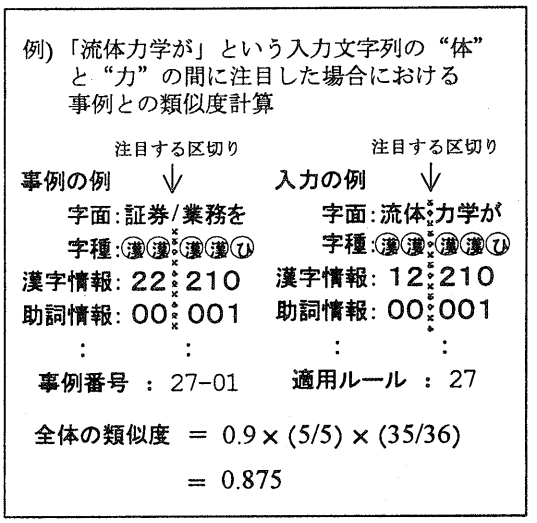


図６：類似度計算の例

見出しとして事例ベースを検索する。ここでは、「証券／業務を」という事例が検索された場合の類似度計算について説明する。まず、字面は一致していないので「字面の類似度」は 0.9 である。また、字種については全ての文字の字種が一致するので一致率は 5/5 となり、「字種の類似度」は 1.0 となる。さらに、その他の表層情報については、属性値が一致する個数を全属性数で割った数を計算する。それぞれ、字種ごとに考慮すべき属性を配列によって与えており、類似度が人間の判断により近くなるよう定義した。例では、考慮すべき属性数は 36 個で、一致する属性は１文字目の漢字情報以外の全てであり 35 個である。したがって、「その他の表層情報の類似度」は 35/36 である。最終的な入力文字列「流体力学が」と事例「証券業務を」の類似度は、全ての類似度を掛け合わせて 0.875 となる。

なお、類似度の最大値は 1.0、最小値は 0.0 となるよう正規化した。

## ３．５　事例の獲得、適用

本手法では、知識を適用した結果、区切られた箇所が誤っている場合に、ユーザからの指摘によって事例を獲得する。獲得した事例は、注

目する文字間で適用されているルール番号別に整理、分類し格納される。なお、重複する事例が事例ベース中に存在する場合、その事例は獲得しない。

次に、入力されたテキストをルールによって分かち書きを行った後、事例を適用する。まず各文字間において適用されたルール番号を、検索のための見出しとして事例ベース内を検索する。そして、入力テキストと各事例を比較して、前記のように類似度を計算する。ここで、類似度が閾値を超えた事例があれば、注目する文字間に事例を適用し分かち書きを修正する。なお、入力と事例の参照回数をできる限り少なくするために、注目する文字間に事例が適用された時点で、その文字間に対する事例ベースの検索は終了し、次の文字間での比較を開始する。

## 4. 評価実験

情報処理関連のテキスト全９章(81758 文字)のうち５章(16954 文字)を対象に、その事例獲得を行ない、残りの４章(64804 文字)を用いて分かち書きの誤り件数と、事例との参照回数を調べた。その結果を図７に示す。

図７において、事例数が０の場合はルールベースのみを用いた場合に相当する。したがって、事例を獲得することによって誤り数が減少しており、ルールベースのみを用いる場合より事例ベースを併用した場合のほうが、分かち書きの誤り件数が減少していることがわかり、知識の自動獲得による精度の向上が確認できた。

しかし、獲得した事例数の増加に伴って、分かち書きの誤り件数は最終的に１２００件程度に収束しており、全ての誤りを分析してみると、その内訳はおおよそ以下のようになった。

・助詞と自立語の競合…3 割
・カタカナ語についての誤り…2 割
・記号等についての誤り…2 割
・その他の誤り…3 割

（その他の誤りの多くは、「２字熟語」「ひらがな書き自立語」「接頭接尾語」等に対する表層解析時のテーブル記載+漏れが原因である。）

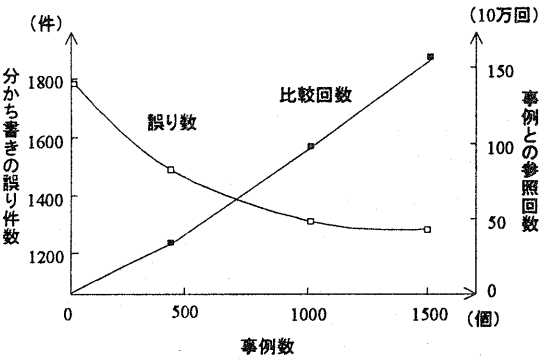

図７：事例数の変化に伴う誤り件数と事例の参照回数の変化

これらの誤りの主な原因は、類似度計算する際の属性情報の不足であると考えられる。このため、不足している属性情報をさらに多く獲得する対策が必要である。また、事例数の増加にほぼ比例して入力テキストと事例の参照回数が増加しており、これに伴い、処理時間も比例して悪化する。このため、実行効率が低下することがわかる。したがって、事例数が増加した際には、精度を維持したまま事例数を減少させる対策が必要である。

## 5. 今後の課題

４章で述べた本手法の問題点への対策について述べる。

**属性情報の不足について**

表層解析に基づく情報のみを用いた場合、1200 件よりさらに誤りを減少させるためには、ルールベースや事例ベースを適用するための情報量が不足している。このため、品詞情報など表層解析では得ることができない情報を付加できる形態素解析を導入することを検討する。また、単語がテーブルに記載されていなかったために表層解析が正しく行われず、分かち書きを誤った例も約２割あったことから、獲得した事例からテーブルに

情報をフィードバックする枠組みも検討している。

**事例数の増加による処理効率低下について**

事例数が増加すると、入力テキストと事例との類似度の計算や、事例獲得時に重複がないかどうかをチェックする際に、事例を参照する回数が増加するため、計算時間が比例して増加し大きな問題となる。このため、獲得した複数の事例から自動的にルールを生成し、分かち書きの精度を維持したまま事例数を減少させる枠組みを研究中である。

## 6. おわりに

本稿では、ルールベースに対して補完的に用いる事例ベースの枠組みについて述べた。さらに、本手法を情報処理関連のテキストに適用した結果、分かち書きの精度が向上したことを確認した。さらに今後の課題について述べた。

本手法ではルールベースに基づく分かち書きが誤った箇所を事例として獲得する。そして、以後の分かち書きにおいて、入力と既存の事例間で表層解析に基づく属性情報を基に類似度を計算し、事例に基づいて分かち書きを修正する。これにより、知識の獲得が自動的に行われ、精度の向上が容易に図れることになった。

現在、更なる精度と処理効率の向上を図るため、音声合成用の形態素解析を導入した枠組みを構築しており、さらに事例からルールへと知識を変換する枠組みについても検討中である。

本研究の出発点は、点訳ボランティアの方々の苦労を少しでも軽減できる、実用的な点訳システムを作ることである。現在はその一部分である分かち書き処理部分についての手法の検討中である。提案する手法をより一般的なものとし、最終的には全ての問題（漢字かな変換、点字表記変換、かな点字変換等）を解決・実装し、実用的なシステムとしてボランティアに配布することを目標としている。

## 参考文献


[1] 日本盲人社会福祉施設協議会点字図書会:点訳の手引き(第2版)(1991).

[2] 日本盲人社会福祉研究会:最新点字表記事典(1990).

[3] 小野,西森,平岡,鈴木,狩野,西原:知識ベースに基づく対話型点字翻訳システム，第54回情処大会,4B-9(1997).

[4] 平岡,西森,小野,鈴木,狩野,西原:知識ベースに基づく点字翻訳のための日本語分かち書き手法，第54回情処大会,4B-8(1997).

[5] 鈴木,小野,平岡,狩野,西原:知識ベースに基づく点字翻訳のための日本語文節区切り手法,言語理解とコミュニケーション研究会,電子情報通信学会,NLC97-28(1997).

[6] 鈴木,小野,平岡,狩野:点字翻訳ボランティアのための対話型分かち書き支援システム,http://csl.sony.co.jp/person/nagao/nlsym97,自然言語処理シンポジウム「実用的な自然言語処理に向けて」(1997).

[7] E.Suzuki,S.Ono,T.Hiraoka,and H.Kanou:Interactive Jaoanese Sentence Segmentation System for Translating Japanese into Braille,Proceedings of the NLPRS97,Vol.1,pp621-624(1997).

[8] 鈴木,小野,狩野:点字翻訳ボランティアのための対話型分かち書き支援システム，自然言語処理学会誌 Vol.5No.4pp.95-110(1998).

[9] 小林重信:事例ベース推論の現状と展望,人工知能学会誌 Vol.7No.4pp.559-566(1992).

[10] 松原仁:推論技術の観点から見た事例に基づく推論，人工知能学会誌 Vol.7No.4pp.567-575(1992).

[11] 高木,小野,鈴木,宮下,狩野,西原:ルールベースと事例ベースに基づく対話型点字翻訳システム，第55回情処大会,3Q-07(1998).