3J−2

# 大語彙かな漢字変換
## —日本語テキストによる文章解析評価—

浅川泰彦[1]　大山裕[2]

日本電気オフィスシステム株式会社[1]
日本電気株式会社 C&C システム研究所[2]

### 1　はじめに

かな漢字変換技術は、世の中に認知された。そして、より精度を上げるために、新しい技術の研究が始まった。しかし、技術を評価する手法が確立されていないため、処理の結果（変換率）に関する議論がほとんど行えなかった。

評価手法の体系がないことは、自然言語処理全般でも同様であり、機械翻訳の翻訳結果に関する評価技術に関する報告 [1] や、開発支援システムに関する報告 [2] があるだけである。

そこで、本稿では、大量の電子化した言語現象（日本語テキスト）を対象にした、コンピュータによる解析処理を生かした評価体系・評価手法を、かな漢字変換を例に提案する。

### 2　言語研究の背景

言語研究の対象は、2つある。[3]

1. 万人に共通な社会的体系の言語 (ラング)
2. 個人の個別的事例の総和の言語 (パロール)

ソシュールが、個々人の言語活動は多種多様で、科学の対象となり得ないと結論して以来、ラングを対象にする言語研究が主流になった。まず、言語の様々な要素を分析し体系を導き出す構造主義が生まれ、次に、言語能力の仕組みを分析しモデルを可能にした変形生成文法が生まれた。

ただし、従来の言語学は、実際に人間が行う言語活動から見れば、わずかな言語データしか扱っていない。そこから演繹的な方法で規則や体系の理論を作成したが、その理論が有効であるかの実証は、誰にも行えなかった。

コンピュータによって、言語の個別的事例の総和を帰納的に分析し、仮説としての体系・規則を作成することが可能になった。そして、自然言語の研究でも、よりやく仮説・実験・検証という科学的な研究方法の必要性が提案 [4] されるようになった。

### 3　評価体系・評価手法の提案

大量の電子化した言語現象を対象とし、大量の辞書データをもとにするコンピュータによる解析処理を用いた実験・検証という評価手法を提案する。

文章解析の評価手順を示した図1をもとに、以下に評価手法の概要を説明する。

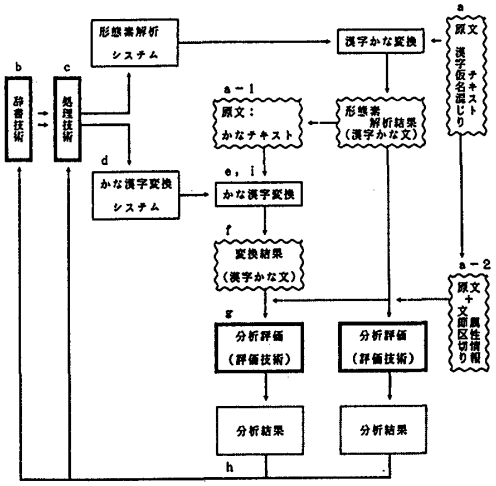


図1　かな漢字変換での文章解析の評価手順

a: 大量の言語現象 (文書・本等) を収集し、電子化テキストを作成する。
b: 大語彙データ (大語彙辞書・様々な制約辞書等) を作成する。
c: 言語現象の処理技術 (文法・ルール・解析アルゴリズム等) を作成する。
d: 辞書や処理技術をもとに自然言語処理システムを構築する。
e: 大量の電子化テキストを、該当する自然言語処理システムで処理する。
f: 処理したデータを、編集可能な形でセーブする。
g: 処理結果を、もとの言語現象と比較し、差異の原因分析を行う。
h: 分析結果をもとに、処理技術の修正・変更や辞書の充実を行う。
i: 見直した処理技術・辞書で再び、テキストを自然言語処理システムで処理する。

---

Kana-kanji translation method using large scale dictionary
—Sentence Analysis Evaluation using Japanese Text—
Yasuhiko ASAKAWA and Yutaka OHYAMA
NEC Office Systems, Ltd.
C&C Systems Research Laboratories, NEC Corporation

## 4 文章解析に使用するテキスト

3で提案した評価体系で使用する、大量の電子化テキストに関して検討した。

### 4.1 テキスト形式の種類

電子化テキストの形式は、以下のように分類できる。

a) 基本テキスト：
- 収集元である、紙または電子のテキストの形式に従った電子化テキストであり、段落や章立てなどの情報も含む。

b) 評価テキスト：
- 1文単位に基本テキストを分解した形式である。(処理しやすいよう 100 文〜 500 文に分割してもよい。)
- 恣意的に抽出した資料として、客観的な評価を得るために使用する。
- 段落は壊すが、文の前後の関係は残す。
- 本テキストから、【翻訳文・かな文・文節区切り文等】を作成する。

c) 検証テキスト：
- 評価テキストから、ある目的の為に、1文単位で再編集したテキスト
- 文の前後関係のつながりは保障しない。
- コーパスとして以下に使用する。
  (a) システムの信頼性評価
  (b) 機能強化項目をチェック
  (c) 辞書データの抽出

### 4.2 テキスト内容の分類

社会生活の中では、様々な言語現象が存在する。自然言語処理の研究を行う上で、言語現象に関するを整理した報告は少ない。[5] そこで、評価に使用するテキストを収集するにあたり、電子文書処理を必要とするある下記のような場面 (位相) を想定して、テキストの分類基準を作成した。資料の収集は、この分類に基づいて行った。

a: 事務 (ビジネス) 関連
　社内文/社外文/法律文書/登記文書/その他
b: 公共 (マスコミや行政) 関連
　政治/国際/経済/スポーツ/科学/文化/社会
c: 教育関連
　文学・言語/社会/数学/理科/その他
d: 日常
　文芸/雑文/宣伝・告示/手紙/記録
e: 専門分野
　学術/辞典・事典/その他

## 5 かな漢字変換評価への適用

本評価手法は、実際に大語彙かな漢字変換 [6] の評価に用いた。

評価に使用するテキストは、4.1の bを採用した。具体的な評価テキストの形式及び変換結果は図 2 の通りであり、変換率及び未登録語含有率は、図 2 の文節区切りテキストと変換結果テキストとの比較により算出する。

◆漢字かな混じりテキスト（図1の原文aに該当）
1:日本とその諸地域。
2:本州・北海道・九州・四国の四つの大きな島のほかに、多くの小さな島からなっています。

◆かな読みテキスト（図1の原文a−1に該当）
1:にっぽんとそのしょちいき。
2:ほんしゅう・ほっかいどう・きゅうしゅう・しこくのよっつのおおきなしまのほかに、おおくのちいさなしまからなっています。

◆文節区切りテキスト（図1の原文a−2に該当）
1:日本と／その／諸地域。／
2:本州／・／北海道／・／九州／・／四国の／四つの／大きな／島の／ほかに、／多くの／小さな／島から／なっています。／

◆変換結果テキスト例（図1の変換結果fに該当）
1:日本とその諸地域。
2:本州・北海道・蝦収・四国の四つのおおきな死麗の外に、多くのちいさな死麗から為っています。

図 2　評価テキストの内容

評価の結果、以下のことが明らかになり、本評価手法が有効であることを確認した。

a) 処理の結果を自然言語処理のレベルで数値化できた。(変換率の向上・未登録語含有率の減少という事実が教育関連の分野全般に表れた。)
b) 比較作業の結果から、新しい処理・辞書情報が得られた。
c) 処理・辞書の強化・修正による影響を、数値として具体化した。

## 6 おわりに

本稿では、大量の電子化した言語現象 (テキスト) を対象とし、大量の辞書データをベースにするコンピュータによる解析処理を用いた実験・検証という評価手法の提案を行い、大語彙かな漢字変換を例に、本手法が有効であることを報告した。

今後、かな漢字変換の評価で、以下の項目を検討したい。

a) 解析失敗例のデータの分析結果をデータベース化する。
b) データベースから、辞書・処理規則へ反映する手順を作成する。
c) 評価作業全体で、人手が介在する部分を少なくする (自動化推進)。

また漢字かな変換や音声出力など他の自然言語処理でも、本評価手法が応用できるかを検討していきたい。

### 参考文献

[1] 長尾, 情報処理 vol.26 No.10,P1197-1201,1985
[2] EDR,“ＥＤＲ電子化辞書”EDR TR-016,1989
[3] Ferdinand de Saussure,“言語学原論”,1928
[4] 横井,“日本語の情報化技術.3”,bit vol.21 No.3,1989
[5] EDR,“概念辞書 (第 2 版) ”P62-63,EDR TR-012,1989
[6] 山田他,“大語彙かな漢字変換” 情処 41 全大,1990