自然言語処理 122－14
（1997．11．21）

曖昧な文節区切りに対応したかな漢字変換評価用テキストデータ

小山　泰男　　安武　満佐子　　吉村　賢治　　首藤公昭
（福岡大学　工学部）

べた書きかな文字列を入力し漢字かな交じり文に変換するかな漢字変換システムの変換精度を評価するテキストデータの場合、入力かな文字列と正しい分かち書き位置を与えておくが、補助用言・形式名詞・拡張表現・接辞・複合語などにおける分かち書き位置は人間にとっても曖昧で、これを一般の自立語と同様に取り扱った場合、分かち書きに対する正確な評価が得られない。そこで、固定的な分かち書き位置と、曖昧な分かち書き位置の２種類を持つ評価用テキストデータを、精度確保のため人手で作成した。本かな漢字変換用テキストデータは、４文節以下の評価用テキスト10,733文、５文節以上の評価用テキスト12,192文を持つ。

The text data to evaluate the accuracy of Kana-to-Kanji conversion with the ambiguity of segmentation.

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo
Fukuoka University
Fukuoka, 814-80 Japan

abstract

In evaluating the accuracy of Kana-to-Kanji conversion, it is required to check whether the segmentation of input Kana string was done successfully. But, there are sometimes arbitrariness in separating auxiliary declinable word, formal noun, extended expression, affix, compound word in the input string. We developed a text data for the evaluation of Kana-to-Kanji conversion accuracy which copes with the arbitrariness of the segmentation. Our text data has two types of specification. One is the obligatory boundary, another is the optional. We have 10,733 sentences including 4 bunsetsus and less, and 12,192 sentences including 5 bunsetsus and more. This text data shall perform the effective evaluation of Kana-to-Kanji conversion.

１．はじめに

べた書きかな文字列を入力し、漢字かな交じり文に変換するかな漢字変換システムの変換精度を評価する場合、特定の少量の評価用テキストデータを人手で入力して変換し人間が目で見て変換精度を評価するか、大量のかな漢字交じり文からべた書きかな文字列を生成して評価用テキストデータとし[3]、主に変換に成功した文節数と漢字数を評価項目としていた。

しかし、最近ではかな漢字変換も共起情報を利用した同音語処理を行うようになるなど、より深く日本語文を解析するようになったため、変換精度も文単位に正しく解析・変換されたかを評価する必要も出てきた。文節分かち書きの失敗やまったく異なる文節の候補表示は文の解析の失敗であるのに対して、補助用言や形式名詞などの取り扱い方の違いや表記のゆれによる文節の候補表示の誤りなどは文の解析の失敗とは言い難い。たとえば、従来のかな漢字変換評価用テキストファイルは単一の文節分かち書き位置しか認めないのに対して、一般に

は「走ることはない」、「走ることは／ない」「走る／ことは／ない」など複数の文節分かち書き位置が考えられる場合が多く、これらを許容した評価は行われていない。

ここでは、文節分かち書き位置が異なるかな漢字変換でも変換精度の比較評価を行えるかな漢字変換評価用テキストデータの作成と文単位の変換精度評価方法を提案する。

2．かな漢字変換システムの誤変換

かな漢字変換システムの誤変換は、下記のように大きく３つに分類できる。以降、例文における「／」は文節分かち書き位置を示す。

1） 文節分かち書きの誤り

最も重大な誤変換の原因であり、通常のかな漢字変換入力システムでは文節分かち書きの修正が必要で、ユーザの操作負荷を増大する。文節分かち書きを誤る原因には次のようなものがある。

・ 辞書に単語が未登録の場合

解析に必要な単語が辞書に登録されていない場合であり、期待される変換は行われない。次の例は「クストー」という単語が辞書に登録されていない場合である。

正解例　クストーは／偉大である
誤変換　クスと／ーは／偉大である

・ アルゴリズム上解析できない場合

辞書には解析に必要な単語が登録されているにも関らず、通常人間の解釈とは異なる文節分かち書きを優先してしまう場合で、個々のかな漢字変換システム（以降、処理系という）により様々なケースが考えられる。次の例は正解例の２文節分が１文節とされてしまった場合である。

正解例　　文章を／こう／書く
誤変換例　文章を／降格

・ 単文内の処理では解決できない場合

かな漢字変換を行うときに入力されたかな文からは、人間の解釈でも複数の文節分かち書きが行われる可能性がある。次の例は、単文内ではどちらが正解とも言えない場合である。

変換例　遠くの／山を／見ている
誤変換　遠く／野山を／見ている

2） 文節内の候補表示順序の間違い

文節分かち書きは成功しているが、変換直後の分かち書きされた一部の文節内の候補表示順序に誤りのある場合で、変換結果の修正作業では文節候補に対して次候補を表示すればよく、操作負荷はあまり増大しない。

・ 意味の違う同音異字語を表示

まったく意味の違う同音異字語を表示してしまう場合であり、次の例では、「読んでいる」とすべきところを「呼んでいる」と誤った変換を行っている。

変換例　彼は／本を／読んでいる
誤変換　彼は／本を／呼んでいる

・ 意味は同じだが表記の違う漢字を表示

単語の意味は同じだが、送り仮名、まぜ書き、かな表記などの表記のゆれによって求める結果と異なる場合であり、次の例では、「他に」と「ほかに」、「良い」と「よい」、「無い」と「ない」がそれぞれ意味は同じだが表記が異なる。

変換例　他に／良い／手段が／無い
誤変換　ほかに／よい／手段が／ない

3） 単語未登録

文節分かち書きは成功しているが、変換直後

の分かち書きされた一部の文節の候補表示に誤りがあり、さらに当該文節候補群の中に求める文節候補がない場合で、変換結果の修正作業においては、単漢字単位に分解して再変換するなど手段を用いるため、修正作業におけるユーザの操作負荷を増大する。次の例は、「香具師」という単語はないが「椰子」「ヤシ」という2つの単語があり、意味は異なるが文節分かち書きには成功した場合である。

変換例　祭りで／物を／売る／香具師
誤変換　祭りで／物を／売る／椰子

3．文節分かち書きの曖昧さ

処理系で用いられる辞書の登録基準と文節分かち書きの基本アルゴリズムの違いは、ここでは取り扱わないが、処理系によって文節に対するの構造定義が異なるため、結果として下記のように文節分かち書き位置が異なる場合がある。

1）補助用言

処理系によって取り扱う補助用言が異なり、補助用言の前で文節を区切るものと区切らないものとあるため、文節分かち書き位置が異なる。

例　この／店で／本を／買って／ください
　　この／店で／本を／買ってください

2）形式名詞

処理系によって取り扱う形式名詞が異なり、形式名詞の前で文節を区切るものと区切らないものとあるため、文節分かち書き位置が異なる。

例　本を／読む／ことが／楽しい。
　　本を／読むことが／楽しい。

3）拡張表現

処理系によって取り扱う付属語が異なり、「において」「に／おいて」などの関係表現や、「に違いない」「に／違いない」などの助述表現に関して文節分かち書き位置が異なる。

例　学校において／授業を／行う
　　学校に／おいて／授業を／行う

　　彼は／楽しいに／違いない
　　彼は／楽しいに違いない

4）接辞

処理系によって、取り扱う接辞が異なり、さらに接辞の前後で文節を区切るものと区切らないものがあるため文節分かち書き位置が異なる。下記の例は、接辞を文節として分かち書きしている。

例　軍縮／論　開発／課　固定／的
　　諸／外国　各／担当／者

5）複合語

処理系によって、辞書登録基準が異なり、次のような複合語は、1単語で登録されていたり、分解して複数の単語で登録されていたりするため、文節分かち書き位置が異なる。下記の例の場合、複数文節に分かち書きしている。

例　大蔵／大臣　国際／空港　非常／食品
　　供給／過剰　過剰／在庫　在宅／勤務

4．かな漢字変換評価用テキスト

かな漢字変換の変換精度を評価する場合、まず文節分かち書きが成功しているかどうかを調べる必要がある。文節分かち書きが成功していることは、各文節の読みと位置で調べることができる。大量の文例を用意するにはコーパスから形態素解析を行い、さらに読みを付ける必要があるが、形態素解析や自動読み付けに精度上問題があり、評価テキストとしての精度を確保するため本論文では人手で作成した。

変換評価の機械処理をより単純化するため、次の例のごとく、べた書き入力文（#）と、文節分かち書きされた期待するかな漢字交じり

文（A）と、それと同様の文節分かち書きを行ったかな文（B）を１対の評価用文として用いるよう収集した。例の行頭につく＃、A、Bは、それぞれの種類を示している。

例　＃にわにはながさく
　　A庭に／花が／咲く
　　Bにわに／はなが／さく

また、処理系における文節構造の基準の違いによる文節分かち書き位置の差異を考慮して、文節分かち書きすべき位置の区切り（／）と、文節分かち書きをしてもしなくてもいい位置の区切り（．）を付加し、文節区切りの比較は以下のように行う。

評価用テキスト	変換結果	評価
／	／	○
．	／	○
．	分かち書きなし	○
分かち書きなし	／	×

従って、次の例の変換評価用文を用いれば、処理系によって変換結果が「咲いている」「咲いて／いる」となるものの、どちらも分かち書き成功とみなすことができる。

例　＃：にわにばらがさいている
　　A：庭に／バラが／咲いて．いる
　　B：にわに／ばらが／さいて．いる

なお、かな漢字変換を評価するにあたり、比較的短い文と長い文では、変換負荷の違いもあるため、期待する確実な文節区切り（「／」）が４個以上、すなわち５文節以上の文と４文節以下の文の２群に分け、　検索が容易になるよう入力かな文字列順にソートした。

４文節以下の文	10,733 文
５文節以上の文	12,192 文
合計	22,925 文

５．かな漢字変換評価システム

自動かな漢字変換評価システムは、ＯＳ（ｵﾍﾟﾚｰｼｮﾝ ｼｽﾃﾑ）上で次の関数が用意されていれば容易に自動かな漢字変換評価システムを作成できる。

1）かな漢字変換システムをオープンする。
2）かな文字列をかな漢字変換システムに入力する。
3）かな漢字変換を行う。
4）文節分かち書き単位にかな漢字文字列を出力する。
5）文節分かち書き単位にかな文字列を出力する。

なお、表記のゆれをチェックするため、処理系は、かな漢字変換用辞書に次のような表記のゆれ情報を持つものとする。

送り仮名	売り上げ／売上げ／売上
長音の有無	メモリー／メモリ
拗音のゆれ	バッファ／バッフア
ヴァとバ	ヴァイオリン／バイオリン
記号	電話／℡
かな	ください／下さい
カナ	ウナギ／鰻
漢字	浜／濱

ここで、出力された、文節単位の文字列により次の評価を行う。説明のため、入力かな文を＃、評価用かな漢字文をA、かな文をB、変換結果としてのかな漢字文をC、かな文をDとする。

1）文節分かち書きされたかな文字列が、文節分かち書きされた評価文字列と一致すれば、分かち書き一致。

＃：おきゅうりょうもあげてほしいですね
B：お.きゅうりょうも/あげて.ほしいですね
D：おきゅうりょうも/あげて/ほしいですね

2）分かち書き一致で、漢字表示が一致しなかった場合、文節単位に期待される文字列の語のゆれを辞書引きによってチェックし、これと一致すれば、ゆれ表示一致。

＃：おきゅうりょうもあげてほしいですね
Ａ：お.給料も/上げて.ほしいですね
Ｃ：お給料も/上げて/欲しいですね

3）分かち書き一致で、漢字表示が一致すれば、完全一致。

＃：おきゅうりょうもあげてほしいですね
Ａ：お.給料も/上げて.ほしいですね
Ｃ：お給料も/上げて/ほしいですね
Ｂ：お.きゅうりょうも/あげて.ほしいですね
Ｄ：おきゅうりょうも/あげて/ほしいですね

なお、今回の評価方法としては、変換直後の結果が正しいかな漢字変換結果であるかを評価することを基準とするため、次の評価は行わない。

1） 文節分かち書きは誤っているが、漢字表示は一致しているもの。
2） 文節分かち書きが成功し、漢字表示が一致しない場合で、漢字表示が一致しない文節の文節候補群の中に期待される文節候補が含まれるもの。

実際に試作したものの処理系の手順は概略次のとおりである。

- 図１の画面で、辞書や入力ファイルを指定する。
- 変換評価処理を実施する。
- 図２の画面のように変換評価結果が表示される。
- 図３のように同時に変換結果やエラーのチェック表が作成され、外部記憶装置に格納できる。

参考までに、市販かな漢字変換システムWXG Ver.2.05の評価は下記の結果となった。

４文節以下の評価用テキストの場合

総行数：10,733　総文節数：38,040

完全一致行	5840(54.4%)
ゆれ許容一致行	7154(66.7%)
分かち書き一致行	9676(90.2%)
一致文節	27717(72.9%)
ゆれ許容一致文節	30180(79.3%)

５文節以上の評価用テキストの場合

総行数：12,192　総文節数：153,450

完全一致行	2504(20.5%)
ゆれ許容一致行	4106(33.7%)
分かち書き一致行	8292(68.0%)
一致文節	95612(62.3%)
ゆれ許容一致文節	101926(66.4%)

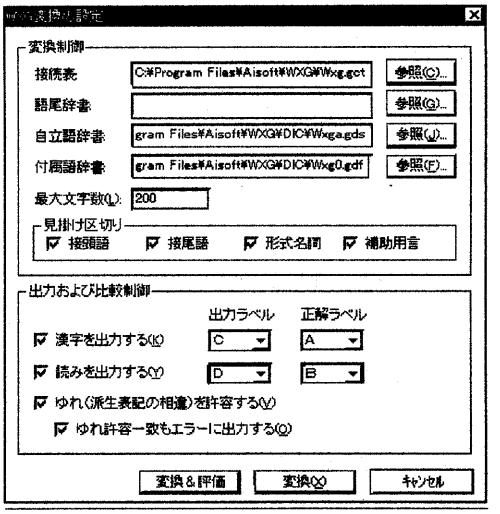

図1　環境設定画面

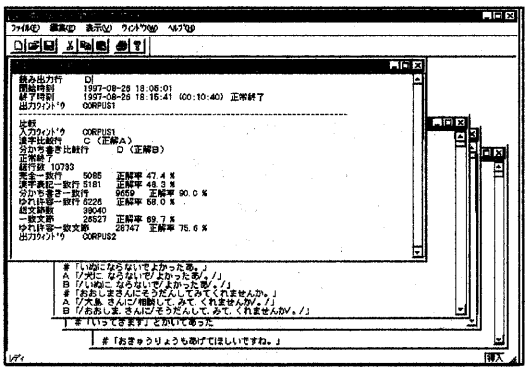

図2　変換評価結果表示画面

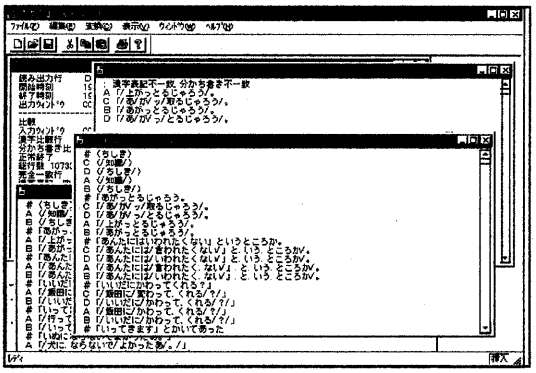

図3　変換結果表示画面

6．おわりに

本かな漢字変換評価用テキストファイルは、市販の文例集や表現集と日経新聞の一部を参考に様々表現を網羅できるよう文例を収集し、読みと文節区切りを人手で付加して作成した。これにより、評価用テキストデータの精度を確保できると共に、比較的容易に自動かな漢字変換評価システムが作成できる。

今後は、大規模なかな漢字変換評価テキストを作成するためには、形態素解析を利用してコーパスから文節分かち書きされた評価文収集を行い、接辞や熟語等区切りの曖昧な部分を人手で補完して読みを付加する変換評価テキスト収集支援システムが考えられる。

また、本論文ではかな漢字変換の評価を、文解析の立場から完全一致、ゆれ表示一致、分かち書き一致の３項目としたが、かなを漢字に変換するという観点で考えれば、文節分かち書きは失敗しているものの漢字表示は一致しているものと文節分かち書きが成功し、漢字表示一致しない場合で、文節候補群の中に期待される文節候補が含まれるものの２項目も補助的に付加した方が良いと考える。

なお、本かな漢字変換評価用テキストの文節分かち書きされたかな漢字交じり文を用いて、かな漢字交じり文を入力とする形態素解析の評価を行うこともできるため、今後はかな漢字変換システムの評価とともに、形態素解析エンジンの評価も行っていきたいと考えている。

参考文献

[1] 首藤,楢原,吉田:日本語の機械処理のための文節構造モデル,電子通信学会論文誌,vol.62-D,NO.12,1979

[2] 吉村,武内,津田,首藤:コスト最小法を用いた日本語文の形態素解析,情報処理学会自然言語研究会資料 60-1（1987）

[3] 清水,橘本,山本:大規模テキストを対象とした仮名漢字変換評価システムの構成と性能評価,情報処理学会自然言語処理研究報告 95-1