5C−3

字面処理による日本文誤り検出の一方式

下村秀樹，酒井貴子，並木美太郎，中川正樹，高橋延匡
(東京農工大学　工学部　電子情報工学科)

1．はじめに

我々は，字面処理あるいは形態素解析に基づいて，日本語文章中の誤りを検出する手法の研究を行っている[1]．一般的に文字コードだけを主な情報とする字面処理に基づく手法は，形態素解析に基づく場合よりも誤り検出能力が低いと考えられる．しかし，（1）高速な処理が可能である，（2）実現が比較的容易である，などの特長もあり，誤り検出能力と併せて評価しなければ，実用性がないと断言することはできない．また評価の結果，実用的でないことが明らかになったとしても，形態素解析やさらに高度な誤り検出手法を評価するための比較データとして，その誤り検出能力を明らかにしておく必要がある．

さて，字面処理で文中の誤りを検出する代表的ツールに，英語の spelling checker がある．これは，空白で区切られた文字列パターン（単語）を辞書と照合し，辞書にないものを誤りであると指摘する．日本語は英語と違い，文中の単語を分かち書きする習慣がない．したがって，単語単位で辞書と照合するこの手法をそのまま応用するためには，字面処理の枠組みを越えた，形態素解析が必要になる．

しかし，誤り検出のために辞書と照合する単位が，単語である必然性はない．そこで我々は，日本文の字種（漢字，平仮名，片仮名，英字，数字等）の変化に着目して文から文字列パターンを切り出し，それを辞書と照合するという方式で，誤りをどの程度検出できるのかを調べた．実験は，パターン切出し規則（後述）を変え，3回行った．

本稿では，3回の実験のそれぞれの誤り検出能力を，第1種の検出誤り（誤りを見逃す），第2種の検出誤り（誤りでないものを検出する）の観点から調べた結果を報告する．

2．誤り検出能力の実験

2.1　実験システムの構成と実験の方法

実験は，図1に示すツールを用い，次の手順で行う．

（1）パターン切出し規則を決め，文章をパターンの並びに分割する，パターン切出しツールを作る．

（2）パターン切出しツールを使って，大量の学習用文章（辞書生成用データ）から，辞書を生成する．

（3）実験用に用意した，誤りを含む評価用文章と誤りのない評価用文章を，（2）で作った辞書と照合し，第1種・第2種の検出誤りを調べる．このとき，出力は人間が見やすいように，KWIC (Key Word In Context) の形式にする．

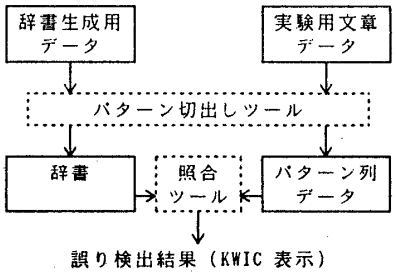

図1　誤り検出実験システム

2.2　実験用文章データの収集

実験のためのデータとして，東京農工大学工学部数理情報工学科の 1989 年度の卒業論文・修士論文から，次のものを集めた．

（1）誤りのない大量の文章データ（辞書生成用データ）

論文 50 編（約 130 万字）を用意した．各論文は人手によってチェックし，誤りはできる限り取り除いた．

（2）第1種の検出誤りを調べるための文章データ

辞書生成用データに含まれないものの中から，誤りを人手によって探して抜き出し，120 文を用意した．このとき，字面処理による誤り検出の能力を考え，誤りは誤字，脱字，仮名漢字変換誤りに限定した．以下，これを「誤りサンプル」と呼ぶ．

（3）第2種の検出誤りを調べるための文章データ

人手によってチェックし，誤りをできるだけ取り除いた論文1編（約4万字）を用意した．以下，これを「正サンプル」と呼ぶ．

なお，これらすべてのデータは，図，表，罫線を削除したものである．

2.3　実験

実験は，さきに述べた手順に従って，パターン切出し規則を変えて3回行った．まず実験1では，単純に字種の変化でパターンを切り出した．実験2では，実験1で漢字1文字のパターン（活用語の語幹など）が大量に切り出されたので，それをなくす補助規則を加えた．また実験3では，それまでの規則では，平仮名の極端に長いパターン（文末に多くある）が多く切り出されていたので，それを短いパターンに分割する補助規則を加えた．

A Method for Detecting Errors in Japanese Sentences Using Textual Analysis
Hideki Shimomura, Takako Sakai, Mitarou Namiki, Masaki Nakagawa and Nobumasa Takahashi
Department of Information Science, Faculty of Technology,
Tokyo University of Agriculture and Technology

実験1の規則

（a）字種の変化点で切る．

字種とは，平仮名，片仮名，漢字，数字，英字である（以下，「字種」と言った場合も同じである）．

実験2の規則

（a）字種の変化点で切る．

（b）規則（a）によって，漢字1文字のパターンと平仮名パターンが連続するところは，一つのパターンに併合する．

実験3の規則

（a）字種の変化点で切る．

（b）次のパターンの前後で切る（助詞，形式名詞に対応）

が，や，を，に，へ，で，も，は，ば，と，

から，て，や，こと，もの，ため，とき，ところ

（c）規則（a）（b）によって，漢字1文字パターンと平仮名が連続するところは，一つのパターンに併合する．

（d）規則（a）（b）によって，平仮名パターンの後に平仮名1文字のパターンが連続するところは，一つのパターンに併合する．

3.4 実験結果

実験結果を，表1に示す．第1種の検出誤りの割合（以下，αと呼ぶ）は，誤りサンプルを辞書と照合した結果について，次式で計算した．

$$\alpha = \frac{\text{（検出できなかった誤りの数）}}{\text{（検出すべき誤りの数）}}$$

第2種の検出誤りの割合（以下，βと呼ぶ）は，正サンプルを辞書と照合した結果について，次式で計算した．

$$\beta = \frac{\text{（検出したパターンの数）}}{\text{（切り出されたパターンの総数）}}$$

また誤りサンプルの中から，各実験で検出できたり，できなかったりした代表的な例を，図2に挙げる．

表1　第1種・第2種の検出誤りの割合α・β

値 ＼ 実験	実験1	実験2	実験3
α (%) (データ数)	35.0 (42/120)	23.3 (28/120)	35.8 (43/120)
β (%) (データ数)	5.2 (844/16098)	7.1 (1011/14259)	4.0 (693/16189)

例1：「…できる用になる．…」
実験1　×　…できる／用／になる／．
実験2　○　…できる／用になる／．
実験3　×　…で／きる／用に／なる／．

例2：「…することを，考てみる．」
実験1　×　…／することを／考／てみる／．
実験2　○　…／することを／考てみる／．
実験3　○　…／する／ことを／考て／みる／．

例3：「…平面図をだけを見ただけでは…」
実験1　○　…平面図／をだけを／見／ただけでは／…
実験2　○　…平面図／をだけを／見ただけでは／…
実験3　×　…平面図／を／だけを／見ただけ／では／…

図2　各実験での誤り検出例（○は検出成功，×は失敗）
（"／"は各実験でのパターンの切れ目）

4．考察

まず，検出誤りの割合α，βを考察する．αが最も良いのは，実験2の23.3%であり，誤りサンプルの75%以上を検出できた．このとき検出できなかったものは28例あったが，このうち10例は，辞書生成データ中に誤りが（取りきれずに）含まれていたことが原因であった．したがって，実験2の規則で，辞書生成データ中の誤りを完全に取り除ければ，

$$\alpha = (28-10)/120 = 13.3\%$$

となり，誤りサンプルの85%以上が検出できる．ただし，実験2の結果は，αが小さい代わりにβが大きくなっているので，このときの規則が最も良いとは言えない．

α，βの関係から見ると，今回の実験では，αを小さくするようなパターン切出し規則の変更は，βを大きくする副作用があった．一般的にもその傾向があると思われる．しかし，実験1と実験3では，αはほぼ同じであるのにβは実験3の方が小さくなっている．この点から考えると，単に字種の変化でパターンを切り出すといった規則に，実験3のような，日本語の特徴を加味した規則を付加することによって，βの増加を抑えてαを減少させられる可能性もある．

次に，各実験間に生じる第1種の検出誤りの差について，図2の例を使って説明する．本手法では，切り出したパターン間の関係の誤りを検出することはできない．したがって，単純に字種の変化点でパターンを切り出した実験1では，図2例1，例2は検出できなかった．これに対し実験2では，漢字1文字と平仮名列を併合する規則によって，この二つのパターンの間の関係を暗に含めていることになる．したがって，図2例1，例2は検出することができた．実験3では，助詞，形式名詞に対応する文字列で平仮名列を分割したので，それまで辞書になかったパターンが，辞書にある複数個のパターンに分割されてしまい，検出できなくなる例があった（図2，例3）．αの値だけを比べると，実験1と実験3はほぼ同じであるが，パターン切出し規則の特徴が異なっているので，検出できる誤りは違う．

5．おわりに

本実験では，字面処理で切り出したパターンを辞書と比較する方法について，パターン切出し規則を変えて3回の実験を行い，その誤り検出能力を定量的に示した．その結果，辞書中の誤りを十分に取り除くことができれば，誤字，脱字，仮名漢字変換誤りの85%程度まで検出できる可能性があることが分かった．今後の課題は次のとおりである．

（1）大量のデータによる実験

（2）パターン切出し規則の変更による能力の向上

（2）他の誤り検出手法（字面処理，形態素解析に基づくものなど）との定量的比較

参考文献

[1]下村秀樹，他：日本語文書作成支援環境の実現に向けて，情報処理学会第32回プログラミングシンポジウム報告集，pp.97-107（1991）