音声言語情報処理　19－15
(1997. 12. 12)

品詞 N-gram 統計情報を用いた日本語文書における誤り検出法について

石場 正大† 竹山 哲夫† 青木 恒夫††
兵藤 安昭††† 池田 尚志†††

†三洋電機株式会社 ハイパーメディア研究所
††中日本自動車短期大学
†††岐阜大学工学部

あらまし 近年、パソコンの普及に伴い、誰もが容易に文書作成する機会を持つようになった。しかし、作成した文書には必ずといっていいほど、入力ミス、編集中のケアレスミスなどによる誤りが存在する。本稿では、日本語文書中の表記誤りのうち、かな漢字変換の選択誤りや誤字・脱字・余字の誤りを検出することを目的に、正文を形態素解析して得られた品詞 N-gram 統計情報とヒューリスティックルールを用いて、日本語文書中の表記誤りを検出する方式について提案し、評価実験を行った。

キーワード 誤り検出、品詞統計情報、日本語文書、自然言語処理、形態素解析、文書処理

An Error Detection Method Using N-gram Statistical Data on Parts of Speech in Japanese text

Masahiro ISHIBA† *Tetsuo TAKEYAMA*† *Tsuneo AOKI*††
Yasuaki HYODO††† *Takashi IKEDA*†††

†Hypermedia Research Center , SANYO Electric Co., Ltd
††Nakanihon Automotive College
†††Faculty of Engineering, Gifu University

Abstract Recent advances in computer technology have enabled a wide range of people to create documents easily. However, when writing documents, we often make mistakes, such as input errors, or careless mistakes when editing. In this paper, we propose an error detection method using N-gram statistical data on parts of speech and heuristic rules. We will perform experiments to examine the effectiveness of this system.

key words Error Detection,Statistical Data on Parts of Speech,Japanese text,Natural Language,Morphological Analysis,Document Processing

1　はじめに

近年、パソコンの利用環境(インフラ)が整備され、またメーカーから文書作成のための様々なツールが提供されてきており、多くの人が、電子的に容易に文書を作成できる環境が整ってきている。

しかし、作成した文書には必ずといっていいほど入力ミスや編集中のケアレスミスによる誤りが存在する。作成した文書の推考支援・校正支援などの処理について、最近研究が盛んに行われているが、日本語ワードプロセッサの中で現れやすい誤り、例えば、かな漢字変換の誤り、表記のゆれ、誤字・脱字等の検出・訂正についての系統的な手法はまだ得られておらず、十分な実用化の段階には至っていない。

英文では、以前からスペル・チェック機能が実用化されているが、日本語ワードプロセッサでは、ユーザが自分の目で文書中に誤りがないかを確かめるというのが現状である。

これは、主として英文は、「分かち書き表現」であるが、日本文は、「べた書き表現」であるため、語の区切りを認識する単語分割、すなわち形態素解析が、英語に比べはるかに複雑であり、文法処理、構文解析といったより複雑なテキスト処理が要求されるためである。

本稿では、文書中の表記誤りのうち、かな漢字変換の選択誤りや入力・編集中のケアレスミスによる誤字・脱字・余字の誤りを検出することを目的に、正文を形態素解析して得られる品詞N-gram統計情報とヒューリスティックルールを用いて、表記誤りを検出するシステムを提案する。次に、上記誤りを含む日本語文書を用いて、本システムの評価実験を行い、最後に考察を述べる。

2　文書中の誤り調査

一般に提出される文書は、何度も読み直しをしている間に、かな漢字変換の誤りや入力ミス、編集中のケアレスミスなどの修正が行われていると考えられるが、それにもかかわらずユーザが見落としてしまう誤りがある。

日本語文書中に現れやすい誤りを調査し、分類を行うために、日常の業務において文書作成の機会を持つ人が、日本語ワードプロセッサで作成した文書に含まれる誤りを、分析した結果を表1に示す。

今回調査対象とした文書は、

1）技術報告書(50頁)
2）週報(200頁)
3）特許明細書(300頁)

であり、文体表現(「です、ます」調、「だ、である」調の文体や漢字レベルなど)や内容表現の誤りを除いた、表記上の誤りに着目し、調査を行った。

表1 文書中の表記誤りの分類

誤りの分類	件数(比率)
かな漢字変換中の選択ミス 例) 与えられたいベント、かな漢似ついて	14件(16.7%)
ケアレスミスによる誤字・脱字・余字 例) 整理をほぼ終わり、見てきましら、 進めることも必要ある、機能ををを	33件(39.3%)
表記のゆれ 例)文節ぎり－文節切り、ソフトウェアーソフトウエア	16件(19.0%)
格助詞の誤り 例)明細書の作成している	3件(3.6%)
その他(送りがな、誤用、括弧未対応など) 例)取り消し－取消、 絶体絶命→絶対絶命	18件(21.4%)
計	84件(100%)

3　形態素解析と品詞 N-gram 統計情報を用いた誤り検出

3.1 システムの概要

本システムの概要を図1に示す。

まず品詞 N-gram データベースとして、新聞記事データ(約 21 万文)を形態素解析し、得られた品詞列に対し、品詞 N-gram 統計処理を行い、そ

の結果を登録しておく。

次にシステムの起動時に、品詞 N-gram データベースの内容を読み込み、パトリシア構造を利用した高速で効率的な検索を行うため、パトリシア形式のツリー上に展開しておく。

そして次に、日本語ワードプロセッサ等で作成したテキスト文書を入力とし、入力された文書を形態素解析システムBreakfastにより形態素解析を行う。Breakfast は、株式会社富士通研究所が開発を行い、Web上で公開している日本語解析ツールである。

次に、この解析によって得られた品詞列パターンから、4-gramの品詞列パターンを抽出し、パトリシアツリーツリー上に展開した品詞 N－gram データベースを検索し、検索結果を、後述する誤り判定基準とヒューリスティックルールに基づき評価を行う。

最後に、この評価結果に基づき、誤り候補リストを作成する。

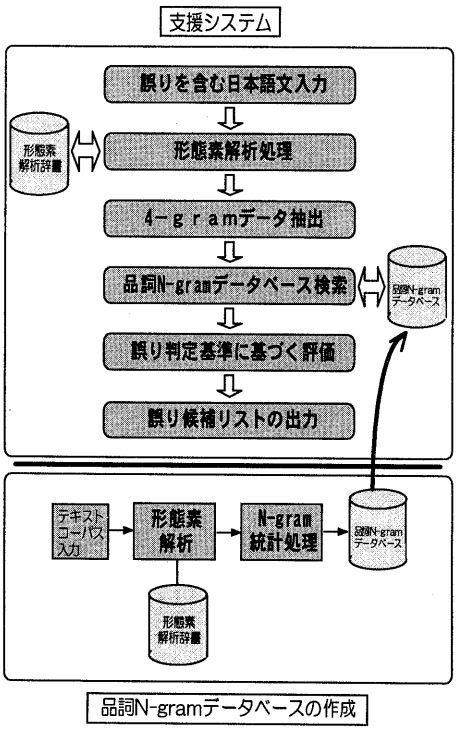

図1 支援システムの概要

3.2 品詞 N－gram データベース

本システムで使用した品詞 N-gram データベースの構築方法について述べる。

まず、朝日新聞記事(1991 年 1 月の記事、朝日新聞社提供)中の社説、天声人語約 21.7 万文を対象に、Breakfast による形態素解析を行い、品詞情報の抽出を行った。

得られた品詞情報(Breakfast では、295の品詞体系に分類)に基づいて、

1) 助動詞、助詞の機能語は、機能語ごとに1品詞として処理
2) 動詞、形容詞は活用形に準拠
3) 普通名詞のうち、1 文字漢字、1文字ひらがな、1文字カタカナは、新たに1品詞として処理
4) 括弧は、開括弧と閉括弧に分割

など品詞の細分化・統合化処理を行い、410個の品詞体系に再分類した(表2)。

次に、形態素解析した結果を、上記方法で分類した品詞体系に基づきパターン化した新聞記事データに対し、4-gram 統計処理を行い、2 回以上出現する品詞4-gram パターンとその頻度を持つデータベースとして構築した。

表 2 使用した品詞体系

品詞	分類	数
名詞	普通、サ変、時相など	5
動詞	未然形、命令形など	53
形容詞	基本形、命令形など	71
助動詞	機能語ごと	143
助詞	機能語ごと	102
副詞	副詞、頻度副詞など	15
特殊名詞	1 字漢字、ひら、カタなど	3
接辞	接頭辞、接尾辞など	10
独立語	接続詞、感動詞	2
その他	未定義語、句読点、括弧	6
	計	410

表3 品詞 N－gram データベース

新聞記事総数	217,179 文
総品詞数	4,922,338 品詞
4－gram 品詞パターン数	275,365 パターン

3.3 誤り判定基準

表記誤りの判定基準として、入力されたテキスト文字列を、Breakfast によって形態素解析した品詞列 $C_1C_2C_3...C_{n-1}C_n$ に対して、

1) 着目する品詞C_mを含む 4-gram 品詞列
($C_{m-3}C_{m-2}C_{m-1}$ C_m 、$C_{m-2}C_{m-1}C_m$ C_{m+1}、
$C_{m-1}C_mC_{m+1}$ C_{m+2} 、$C_mC_{m+1}C_{m+2}$ C_{m+3})
に対し、それぞれ品詞 N-gram データベースを検索し、その品詞列の頻度が0、つまりデータベースに存在しない品詞列が2つ以上存在する場合

2) 着目する品詞C_mを含む上記4つの品詞列の頻度をそれぞれ ω_{m1}、ω_{m2}、ω_{m3}、ω_{m3} とすると、

$$\gamma = \frac{\omega_{m1} + \omega_{m2} + \omega_{m3} + \omega_{m3}}{4} <= \gamma$$

である(γ=2 を採用)場合
を誤り候補とする。

さらにこの判定基準に加え、ヒューリスティックなルールとして、上記品詞列 $C_1C_2C_3...C_{n-1}C_n$から抽出した 4-gram 品詞列中に、

1)1 文字漢字名詞が2文字以上存在する場合
2)1 文字ひらがな名詞が存在する場合
3)1 文字カタカナ名詞が存在する場合
4)未定義語が1つ以上存在する場合
を、誤り判定基準として実験を行った。

4. 評価実験

4.1 組み込んだ誤り例

実験用に埋め込んだ表記上の誤り例と()内に正解を示す。

1)付属語の脱字、余字
例)実際にあた(実際にあった)、計算機活用して(計算機を活用して)、計算のに(計算に)、知らなでしょう(知らないでしょう)、ほとん(ほとんど)
原因)編集中のケアレスミス等に起因
2)誤字
例)断圧(弾圧)、人間性(人間性)
原因)イメージや形の類似に起因
3)キー入力誤り(r,t の押し間違いなど)
例)見てきましら(見てきました)、〜までとすつ(〜までとする)
4)かな漢字変換の選択誤り(フォーカス誤り)
例)ビジュアル成し量(ビジュアルな資料)、基準似ついて(基準について)
5)かなづかい表記誤り
例)1人づつ(1人ずつ)
6)カタカナ表記誤り
例)チェックシテ(チェックして)、オブジェット(オブジェクト)
7)当て字
例)無理矢理(無理やり)

今回の実験では、品詞レベルの統計情報による誤り検出を目的としているため、以下の誤りは対象としていない。

1)同音異義語の誤り
例)異動する/移動する
2)誤用
例)意味慎重(意味深長)、絶対絶命(絶体絶命)
3)カタカナ表記のゆれ
例)アプリケーション−アプリケイション、インターフェース−インタフェース
4)格助詞(てにをは..)の誤り
例)私が学校と行く、汽車で東京が行く

これらの誤りを検出するためには、

1)意味解析処理が必要
2)あらかじめ誤りやすい表記を辞書に登録
3)表記のゆれ辞書や長音の誤り規則が必要
4)構文解析処理が必要
などの処理が要求されるため、対象外とした。

4.2 評価実験

誤りを含むデータとして、品詞 N-gram データベースの元となった朝日新聞記事データより無作為に抽出した121文に対し、4.1 で述べた誤りを29個埋め込んだ場合(close テキスト)と CD-毎日新聞'94年版より抽出した130文に対し、同様の誤りを28個埋め込んだ場合(openテキスト)について、誤り検出評価実験を行った。

5. 実験結果と考察

5.1 実験結果

Breakfast が分割した単語(品詞)を単位とし、適合率(precision)、再現率(recall)を算出した。

$$\text{適合率} = \frac{\text{誤り単語を正しく検出した数}}{\text{誤り単語と指摘した総数}}$$

$$\text{再現率} = \frac{\text{誤り単語を正しく検出した数}}{\text{埋め込んだ誤り単語数}}$$

3.3 で示した判定基準を用いると、埋め込んだ誤りによって、図2に示すように周辺の品詞列の頻度にも影響を及ぼす可能性があるため、誤りを連続して指摘した場合は、それらを1つとしてカウントする。

```
                    36    37    38  39  40      41
【82】...../無理/やり/車/で/に/乗せて/..
        品詞列          頻度
      AgAgAcAh          140
      AgAcAhAk            0
      AcAhAkCx            0
        37単語目でエラーの可能性あり
      AhAkCxAe            0
        38単語目でエラーの可能性あり
      AkCxAeAk            7
        39単語目でエラーの可能性あり
                ⋮
```

図2 連続する誤り検出の例

上記方法により求めた結果を示す(表3)。

表3 誤り検出実験結果

評価値＼対象	朝日新聞 (close)	毎日新聞 (open)
再現率	$\frac{19}{29}$ = 65.5%	$\frac{22}{28}$ = 78.6%
適合率	$\frac{19}{97}$ = 19.6%	$\frac{22}{129}$ = 17.1%

5.2 検出した誤りの分析

実験結果より、正しく検出した誤り、見落とした誤り、誤検出した例とその分析を行う。

5.2.1 正しく検出した誤り

. . . 19件(openでは22件)

1) 品詞統計情報により検出した誤り

. . . 7件(openでは10件)

例)計算/の/に/必要(計算に必要)
法規/で/の/で(法規ですので)
困難な/困難である/こと(困難であること)
の/基準/似ついて(の基準について)
機能/など/を/を(機能などを)

2) ヒューリスティックルールにより検出した誤り

. . . 12件(openでは12件)

例)実際/にあた/ケース(実際にあったケース)
知/ら/な/で/しょう(知らないでしょう)
チェックシテ/いる(チェックしている)
1人/づ/つ(1人ずつ)
と/す/つ/こ/と/は(とすることは)

close、openテキストでの、誤り検出数の差は、
・国民/を/断/圧し/、⇔者/を/断/圧した/翼賛
・交代/が/貧/パン⇔以前/より/貧/パン
・デザイナー/たちがそ/美/しさ⇔
安/さや/そ/美/しさ
(左:close、右:open、右のみを誤りとして検出)
である。これは活用語尾や誤りを入れた位置で形態素解析した品詞列パターンの違いによるものである。これらを見落とした原因については以下で述べる。

5.2.2 見落とした誤り

. . . 10件(openでは6件)

1) 誤り検出用の形態素解析システムを必要とするもの . . . 2件

例)共有/しいる/から(共有しているから)
計算機/活用/して(計算機を活用して)

現状の形態素解析システムは、文法的に正しい文章を対象としているため、上記例の「名詞」+「動詞」のような、誤った単語列でも正解と判定とする。このような接続を誤りとする誤り検出用に特化した形態素解析システムが必要となる。

2) 構文解析レベルの誤り検出を必要とするもの . . . 1件

例)選択/する/選択/した(選択した)

文法上は問題がないため、構文解析レベルで、係り受けなどの誤り判定が必要となる。

3) 入力時点で検出を必要とするもの
. . . 3 件

例)が/貧/パン/に(が頻繁に)
を/断/圧し/、(を弾圧し、)
受ける/野/ではな/いか(受けるのではないか)

4) 解析システムにおける改良を要するもの
. . . 3 件

例)デザイナー/たちがそ/美/しさ/(デザイナーたちがその美しさ)
無理矢理/引き抜かれ(無理やり引き抜かれ)

5.2.3 正解を誤検出した例
. . . 78 件(open では 106 件)

close、open テキスト共に、ヒューリスティックルールによるものが約 40%、データベースのチューニング問題によるものが約 50%であった。

1) 解析辞書の不備による解析誤りによる

□ ヒューリスティック(未定義語)により誤認
. . . 18 件(open では 20 件)

例)最高/の/3/%/強
ゴーサイン/を/出した
輸銀法/の/改正

□ ヒューリスティック(1文字漢字連続)により誤認
. . . 7 件(open では 14 件)

例)2/国/間

□ ヒューリスティック(1字ひら、カタ)により誤認
. . . 6件(open では 5 件)

例)150/ドル/たら/ず
6/カ/国会/談

2) 解析規則の不備による解析誤りによる
. . . 8 件(open では 5 件)

例)花/にな/って/しまう
当然/の/必要/であるが

3) 品詞 N-gram データベース訓練データのチューニング問題による
. . . 38 件(open では 60 件)

例)支持/さ/れる/ようであって/ほしい
、/大/賛成/である

6. まとめ

日本語ワードプロセッサを使ったことにより、従来手書きでは起こり得なかった入力ミスや編集中のケアレスミスによる誤字・脱字・余字などの表記上の誤りを検出し、指摘することは、ユーザが文書の校正にかける負担を軽減し、有効である。今回既存の形態素解析システムに品詞 N-gram データベースとヒューリスティックルールを用いた表記誤り検出方法について、評価実験を行った。

実験結果より、再現率は、解析システムを整備し、誤り解析用にチューニングすれば、概ね 15%は向上する可能性がある。また適合率も、辞書や規則の整備により、30%前後にまで向上する見通しがある。

さらに今後は、

1) 誤り検出を目的とする形態素解析用接続規則の構築
2) 未定義語に対応する辞書の強化
3) 品詞 N-gram データベース訓練データのチューニング
4) さらに大規模なデータに対する実験

を行う予定である。

謝辞

日頃より本研究に関しご理解いただき、またご指導いただく通信システム研究部 富田部長、細川室長に感謝致します。

参考文献

[1] 鈴木、武田：日本語文書校正支援システムの設計と評価、情報処理学会論文誌、Vol.30,No.11(1989)

[2] 池原、小原、高木：文書校正支援システムにおける自然言語処理、情報処理学会誌、vol.34,No.10,pp1249-1258(1993)

[3] 特集誤った日本語に気付き始めたワープロ：日経バイト 1995 年 January

[4] 颯々野、斎藤、松井：アプリケーションのための日本語形態素解析システム、言語処理学会第 3 回年次大会発表論文集(1997 年 3 月)