情報学基礎 57－7
自然言語処理 136－7
（2000．3．21）

決定リスト，用例ベース手法を用いたコーパス誤り検出・誤り訂正

村田 真樹　内山 将夫　内元 清貴　馬 青　井佐原 均

郵政省 通信総合研究所 関西先端研究センター 知的機能研究室

〒651-2492 神戸市西区岩岡町岩岡 588-2

TEL:078-969-2181 FAX:078-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,mutiyama,uchimoto,qma,isahara}@crl.go.jp

あらまし

近年，種々のタグつきコーパスが作成され，タグつきコーパスを利用した研究も盛んに行なわれている．しかし，タグつきコーパスには多くの誤りが含まれており，このことが各研究の進捗の妨げとなることも多い．そこで本研究では，決定リストおよび用例ベース手法により，コーパスの誤り検出および訂正を試みた．

キーワード　コーパス中の誤り，検出，訂正，決定リスト，用例ベース

Corpus Error Detection and Correction Using the Decision-List and Example-Based Methods

Masaki Murata　Masao Utiyama　Kiyotaka Uchimoto　Qing Ma　Hitoshi Isahara

Intelligent Processing Section, Kansai Advanced Research Center,

Communications Research Laboratory, Ministry of Posts and Telecommunications

588-2, Iwaoka, Nishi-ku, Kobe, 651-2492, Japan

TEL:+81-78-969-2181 FAX:+81-78-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,mutiyama,uchimoto,qma,isahara}@crl.go.jp

Abstract

Various types of tagged corpora have recently been constructed, and various research using the tagged corpora have been done. However, current tagged corpora include many errors, which make trouble for research. Therefore, we tried to detect and correct errors in the corpora by using the decision-list or example-based methods.

key words　Errors in Corpora, Detection, Correction, Decision List, Example-based

1 はじめに

近年，さまざまなコーパスが作られ，教師あり機械学習の研究[1][2][3]をはじめとしてコーパスを用いた多種多様な研究[4][5]が数多くなされている．しかし，コーパスには誤りがつきもので，この誤りが各研究の進捗を妨げる場合も多い[1]．このため，コーパス中の誤りを検出・修正することは非常に重要なことである．

このコーパス中の誤りを検出する試みが最近いくつかなされ始めている[6][7]．まず，内山の研究[6]では，形態素コーパスでの過分割の誤り，例えば，「休憩室」を「休」「憩室」と分割してしまう誤りを検出する方法を提案している[2]．そこでは，「分割した場合の確率」と「つなげた場合の確率」をコーパスより求め，「つなげた場合の確率」の方が圧倒的に大きい場合，分割するのは間違いだと判定する．また，乾らの研究[7]では，構文情報のコーパスでの係り先の誤りを検出する方法を提案している．コーパス中のある文節Xの係り先Yがあっているかどうかを調べる場合，コーパスよりその文節Xがその係り先Yになる確率を求め，その確率が極端に小さい場合その係り先Yは間違いだと判定する．両者の研究は，一般化して考えると，ほぼ同様なことをしており，コーパスのタグがあっている確率と間違っている確率をもとめ，間違っている確率の方が圧倒的に大きい場合そのコーパスのタグを誤りとするという方法である．「間違っている確率」の大きいものを間違っているものと考えるのは自然なことであり，ほとんどのコーパス修正の研究でこの種の考え方を利用することが可能だと思われる．

しかし，先の二つの研究で用いられた手法はいずれも形態素の過分割，係り受け誤りと，それぞれその問題に特化した方法を用いて誤り検出を行なっていたため，その手法の汎用性を見えにくくしている．例えば，内山の過分割の研究では，過分割の検出に特化したような式(例：$P(\text{休憩室})/(P(\text{休})P(\text{憩室}))$[3]が用いられている．また，乾らの研究では，すでにできあがった構文解析システムが出す誤り確率を利用している．この構文解析システムでは，構文解析に特化した情報をたくさんに利用していると思われるし，また，誤りを検出する対象とするコーパス以外の情報を用いている可能性も高く，汎用的なコーパス修正とはいいにくい．

これらの適用範囲の限定された方法を用いる研究に対して，本研究では，一般的にどの問題に対しても用いることができると考えられている用例ベース手法や決定リスト手法を利用し，(辞書情報など他の情報を一切用いず)対象とするコーパスのみから「間違っている確率」を算出し誤りを検出することを試みる．この一般にどの問題に対しても用いることができそうな手法を用いて，もし高精度の誤り修正を実現できると，この手法によりほとんどのコーパス修正の問題を対処できるのではないかと期待される．

また，検出精度の面でも，先の二つの研究はそれほど高い値を出しているものではなかった．内山の過分割の研究では上位150個を出力してそのうち50個程度を正しく誤りと検出する程度で上位での誤り検出率が30%程度とみられまだ実用段階に入ったとはいいにくい．また，乾らの構文誤りの検出の研究でも，上位精度が明確に記述されていないが，1,200個程度調べて50個程度の誤りを検出する程度で，これはまたさらに実用性に乏しいものとなっている．これに対し，本研究では，内山も乾らもまだ対象としていない，形態素情報の誤り検出(品詞・活用型などの誤り検出)を試みたところ，上位での誤り検出率は70%～80%と高く，また4,000個のデータを調べれば2,000もの誤りを検出できそうな見通しも得ている．さらに先の二つの研究ではまだ対象とされていなかった，誤り修正を行なったところ，上位での誤り修正率は65%～80%と高く，誤り検出のみならず誤り修正も比較的容易にできそうな見通しをえている．これらの結果は，問題領域が異なるためによかっただけとも思われるが，形態素情報は言語処理で基礎的で重要な情報であるとされているので，その形態素情報を高い精度で修正可能という結果は十分有用なものといえる．

以降，次節でコーパス修正の考え方を理論的に整理し，3節で，決定リストや用例ベース手法を用いた形態素情報の修正について述べる．また，本稿では，本手法の汎用性を確かめるために，決定リストを用いた構文情報の修正も行なっており，これを4節で述べる．本稿での誤りの検出・修正の対象とするコーパスとしては，京大コーパス[9][10]を用いる．

2 コーパス修正の基本的考え方

2.1 コーパス修正のための評価式

コーパス修正の研究は，このタグは正解，また，このタグは誤りというものがふられたデータがないため，基本的に教師なし学習の問題となる．このため，コーパス修正にはなんらかの基準が必要となる．

先にのべた二つの先行研究では，一般化して考える[4]と，以下の評価基準を利用して，この教師なし問題

[1] 例えば，われわれは排反な規則，つまり，学習データで100%正しい規則を利用する手法で高精度の文節まとめあげを実現している[3]が，コーパス自体間違っている場合は学習データに誤りがあることとなり，学習データで100%正しい規則というのは怪しくなってくるし，また，そういう規則をうまく抽出できなくなる可能性があって，排反な規則の利用自体が困難になる．すなわち，排反な規則を利用したいという立場ではコーパス中の誤りは非常に致命的なものとなっている．
また，逆に規則の作成中に排反な規則が得られない場合，これはコーパス中に誤りがあるためではないかと推測される．このことが本研究の決定リストを用いたコーパス修正の動機付けとなっている．

[2] 単語分割の問題は，情報検索において重要な問題として位置づけられている[8]．

[3] P(休),P(憩室)の部分は，「休」「憩室」の単純な出現率を用いているが，厳密には「休憩室」という文字列が「休」と「憩室」に分割される確率を用いるもので，近似をすでに使ったものとなっている．この近似はデータスパースネスに対処するためのものだが，この近似自体は，過分割の検出と同じような問題でしか使えない．

[4] 両文献ともに，ここで行なっているような一般化した記述や議論をしていない．

を解いていることになっている．

1. 内山の方法

$$\text{評価式} = \frac{\text{修正後のタグが正しい確率}}{\text{修正前のタグが正しい確率}} \quad (1)$$

2. 乾らの方法

$$\text{評価式} = \text{修正前のタグが誤っている確率} \quad (2)$$

これらの値が大きい場合はタグが誤っている可能性が高いと考えるわけである[5]．

内山の式は修正するとどのくらい確率値が上昇するかを示しているものであって，誤り検出のみならず，誤り修正のことも考えた式になっている．検出精度だけでなく，修正精度も重視する場合は内山の方法を採用するとよい．乾らの式は，純粋にタグが誤っている確率を用いるもので，誤りの検出精度を重視する場合に採用するとよい．

本稿では，コーパス修正の問題が人手コストとして修正箇所を見つけることの方が修正よりも難しいことから，誤り修正を重視せずに誤り検出を重視することとし，まず乾らの方法の「タグが誤っている確率」を利用して誤り検出を行なうことにする．また，修正は，検出された箇所において，内山の手法の分子の「修正後のタグが正しい確率」をもとめ，この値が大きいタグを利用して行なう．まとめると，本手法の評価式は以下のようになる．

1. 検出

$$\text{評価式} = \text{修正前のタグが誤っている確率} \quad (3)$$

2. 修正

$$\text{評価式} = \text{修正後のタグが正しい確率} \quad (4)$$

本稿の手法は，内山の研究で示唆されている式の分子と分母を分割して，逐次的に用いているものともいえる[6]．

2.2 確率値の算出方法

「修正前のタグが誤っている確率」や「修正後のタグが正しい確率」といっても，これをどのようにして簡単に求めるかが次の課題となる．ここでは，まず「修正前のタグが誤っている確率」の算出方法を例を使って考えてみよう．

京大コーパスでは，「、」の形態素情報の統計をとってみると，表 1のような結果が得られる．この表は，ちょっと見ただけでも「特殊 読点」となっているデータが圧倒的に大きく他は誤っているということが予想される．ここで「、 、 * * *」の誤りの確率を考えてみる．まず，これの正解確率は，その出現数を総数で割った 3/(3+26540) と考えてよいだろう．誤り確率は 1 から正解確率をひいたものと考えてよさそうなので，1 - 3/(26540+3+2+1) となると思われる．本稿での誤り確率のもとめ方は基本的にこの方法で行なう．

表 1:「、」の形態素情報の統計

、 、 * 特殊 読点 * *	26,540
、 、 、 * *	3
、 未定義語 * 名詞 サ変名詞 * *	2
、 、 * 特殊 記号 * *	1

表 2:「の」の形態素情報の統計

の の * 助詞 接続助詞 * *	25,739
の の * 助詞 格助詞 * *	1,601
の の * 名詞 形式名詞 * *	1,350
の の のだ 助動詞 * ナ形容詞 語幹	398
の の だ 判定詞 * 判定詞 ダ列特殊連体形	191
の の の * *	1
の の * 名詞 普通名詞 * *	1
の の * 連体詞 * * *	1

しかし，単にこれだけでは確率の求め方としてあらっぽすぎる．例えば，「の」の形態素情報の統計をとってみると，表 2のような結果が得られる．ここで，頻度 191 の「の の だ 判定詞 * 判定詞 ダ列特殊連体形」の誤り確率を求めると 99.3% (= 1 - 191/(25739+1601+...)) となってほとんど誤っていると判定される．「の の だ 判定詞 * 判定詞 ダ列特殊連体形」が正しい場合も数多くあり正しいのに，この単純な方法ではこれを全部誤っていると推定してしまう．

そこで，本稿では確率値の算出に用例ベース手法[3] や決定リスト手法[12][13] を利用する．用例ベース手法は，いま解きたいものと良く似た用例を集め，その用例集合での出現率を確率値とする手法である．「のような」の場合「の」は 84 個あってすべて「の の だ 判定詞 * 判定詞 ダ列特殊連体形」であるので，正解確率 100%，誤り確率 0% となり，これを間違って誤りと検出することがなくなる[7]．一方，決定リスト手法は，多くの素性に展開し各素性の確信度を求め，確信度の最も高い素性 (パターン) のときの，正解確率と誤り確率を用いる方法である．先の「の」の例だと，「の」「のような」「名詞＋の」「の＋助動詞」などといろいろなパターンでの確率を求める[8]．この結果を京大コーパスを用いて計算すると表 3のようになる．表の「判定詞の場合の数」は，京大コーパスで各素性に適合する事例における「の」が判定詞の

[5] クラスが二つしかない問題の場合は，上記の二つの基準は等価となる．また，乾らの方法は，内山の方法で分子の値を用いない場合に相当する．

[6] 修正後のタグが正しい確率が大きくなるように修正する式 (1) や式 (4) は，修正後の一文の確率が最大になるように修正を行なう OCR 修正などの研究[11] とも似た考え方となっている．しかし，OCR の研究は，正解情報を意味する大量の生のデータを利用できるために基本的に教師あり学習の研究であり，教師なし学習の研究であるコーパス修正の研究とは大きく異なる．

[7] 用例ベース的確率算出方法は，バックオフによる確率推定を極端なまで行なったことに相当する．また，誤り修正の場合，自分自身だけの事例を用いると一つも誤りを検出できなくなるので，最低自分以外に一つ，合計二つ以上の事例をもってくる必要があり，k 近傍法[14] の k が 2 以上のときに相当する．

[8] ただし，総数が 1 の素性は用いない．

表 3: 「名詞＋のような」の場合の決定リストによる確率値の算出方法

各素性	確信度	正解確率	誤り確率	判定詞の場合の数	総数
のような	100%	100%	0%	84 個	84 個
の＋助動詞	99.5%	99.5%	0.5%	187 個	188 個
の	99.3%	0.7%	99.3%	191 個	29,282 個
名詞＋の	99.2%	0.8%	99.2%	162 個	20,220 個
.....					

場合の数で，「総数」は京大コーパスで各素性に適合する事例の総数である．例えば，「のような」のパターンは，判定詞の「の」だけが 84 個出現したことを意味し，「の＋助動詞」のパターンでは，判定詞の「の」が 187 個，それ以外の「の」が 1 個出現したことを意味する．このデータからの正解確率，誤り確率の求め方は，先にのべたのと同じで，187/188, 1-(187/188) などの計算をして求める．また，確信度，これはその規則の確らしさを意味するが，この確信度としては，正解確率と誤り確率のうち大きい方の値が用いられる．例えば，一つ目の「のような」は確信度 100% でほぼ正しい情報と推測されることになる[9]．決定リストではこの表の最上位にある，この規則を用いることになり，誤り確率は 0 となって，用例ベースと同じく「のような」の「の」は判定詞で正しいと推定され，間違って誤りと推定することはない．上の二行の情報がないときは，誤り確率 99.3%，確信度 99.3% で誤っていると判定される．

次に「修正後のタグが正しい確率」の求め方だが，これは，表 1の読点の簡単な場合で考えると，「修正後のタグ」は頻度の最も大きい「、 、 * 特殊 読点 * *」とすればよく，これが正しい確率はこれの出現数を総数で割ったもの，すなわち，99.99% (=26540/26543) となる．これは単純な場合だが，用例ベース手法，決定リスト手法の場合ともに，誤り確率などを求めた事例集合でこの計算をして「修正後のタグが正しい確率」を求めればよい．

本稿ではおおよそ上記のような方法で確率値を算出する[10]．ここであげた手法は，教師あり学習の手法として様々な問題で用いられているもので，別に形態素情報の修正に限ったことではなく，様々な種類のコーパス修正に容易に利用できる．

3 形態素情報の修正

本節では，形態素情報のコーパス修正を試みた結果について記述する．

[9] この規則は文献(3) でいう排反な規則に相当する．

[10] 本稿では簡単のためにここで述べたような手法で確率値を算出するが，最大エントロピー法(2) やその他の強力な手法で確率値を求めてもよい．

表 4: 形態素情報

	読み情報あり	読み情報なし
全形態素数	487,691	487,691
曖昧形態素数 (のべ)	275,291	270,534
曖昧形態素数 (異なり)	5,957	5,539

3.1 形態素情報の予備調査

まず，対象とする京大コーパスでの形態素情報の調査を行なった．この結果を表 4に示す．表の全形態素数はコーパスにあったすべての形態素の数を意味する．また，曖昧形態素数はコーパスにあった形態素のうち，コーパス中の他の形態素と表記が同じであった形態素の数を意味する．たとえば，「の の * 助詞 格助詞 * *」，「の の * 助詞 接続助詞 * *」といったものは，表記が同じ「の」で異なる形態素なので曖昧形態素と考える．また，この調査では 5 つまでの形態素連続までは「では」と「で｜は」のように形態素の区切りが異なるものが他にある場合も曖昧形態素と考えている．(つまり，この場合，「では」「で」「は」はそれぞれ曖昧形態素となる．) 表中の「読み情報あり」と「読み情報なし」は，京大コーパスが読み情報に弱いという理由から設定したもので，「読み情報あり」は読み情報も含めて曖昧形態素の数を数えたもので，「読み情報なし」は読み情報を省いて曖昧形態素の数を数えたものを意味する．(全形態素数は「読み情報あり」と「読み情報なし」で変わることはない．) 例えば，「読み情報なし」では「日 ひ * 名詞 時相名詞 * *」と「日 び * 名詞 時相名詞 * *」のように読み情報のみが異なる場合，これらを異なる形態素として扱わない．

表からわかるように，京大コーパス約 2 万文には，487,691 形態素が存在しており，人手で 50 万の形態素を徹底的に調べあげるとコーパス修正ができるがそれは非常に大変である．また，曖昧形態素数は，読み情報の修正をあきらめたとしても，270,534 形態素存在しており，修正範囲を曖昧な形態素にしぼったところで網羅的に人手で修正するのは困難である．曖昧形態素数の異なりは，5,539 であるので，曖昧形態素の種類ごとにまとめて出力させ[11]，それを見て人手で修正することも可能かとも思われるが，各種類ごとに多数の事例が出力されると思われ，それを用いた修正も若干無理があると思われる．

上記の議論により，コーパス修正は難しい問題であることがわかる．このため，このコーパス修正を容易に行なう技術を確立することは重要である．

3.2 形態素情報の修正実験

次に，本稿での形態素情報の修正実験について述べる．本稿での形態素情報の修正では，読み情報は対象か

[11] ツリーバングの構築の際に類似した部分を一覧して表示する研究を行なっているもの(15) もある．また，表 1～表 3にあげたようなデータを表示させて，コーパス修正の支援を図ることも考えられる．

表 5: 形態素情報の修正結果

	用例ベース手法 (抽出総数 591 個)			決定リスト手法 (抽出総数 4,054 個)		
	検出精度	修正精度	不明	検出精度	修正精度	不明
ランダム 300 個	43%(130/300)	41%(123/300)	38 個	53%(160/300)	49%(146/300)	17 個
上位 50 個	60% (30/50)	58% (29/50)	0 個	82% (41/50)	78% (39/50)	0 個
上位 100 個	55% (55/100)	52% (52/100)	6 個	70% (70/100)	67% (67/100)	0 個
上位 150 個	49% (74/150)	47% (70/150)	9 個	71%(107/150)	66% (99/150)	2 個
上位 200 個	46% (91/200)	44% (87/200)	17 個	74%(147/200)	69%(137/200)	2 個
上位 250 個	44%(110/250)	41%(103/250)	23 個	78%(194/250)	73%(183/250)	3 個
上位 300 個	42%(125/300)	39%(117/300)	31 個	76%(229/300)	71%(212/300)	4 個

表 6: 形態素誤りの検出修正の基準

該当箇所	修正前	修正後
検出・修正ともに成功と判断したもの		
ロシア側は	ロシア * 名詞 普通名詞 * *	ロシア * 名詞 地名 * *
世界がアッと驚く	アッ * 名詞 普通名詞 * *	アッ * 感動詞 * * *
今日のねじ曲がりを	の * 助詞 格助詞 * *	の * 助詞 接続助詞 * *
検出のみ成功と判断したもの		
戦車五十両を破壊した	両 * 名詞 普通名詞 * *	両 * 接頭辞 名詞接頭辞 * *
「だからこそ国際	だ * 名詞 普通名詞 * *	だ だ 判定詞 * 判定詞 基本形
友達もたくさんいた」と話して	いた * 副詞 * * *	いた いる 接尾辞 動詞性接尾辞 母音動詞 タ形
検出失敗と判断したもの		
アジア・太平洋経済協力会議に	太平洋経済協力会議 * 名詞 組織名 * *	太平洋経済協力会議 * 名詞 普通名詞 * *
国内外の、	国 * 名詞 普通名詞 * * | 内外 * 名詞 普通名詞 * *	国内 * 名詞 普通名詞 * * | 外 * 名詞 普通名詞 * *
を決めるだけで、	で だ 判定詞 * 判定詞 ダ列タ系連用テ形	で * 助詞 格助詞 * *
トップクラスの輸送量	トップクラス * 名詞 普通名詞 * *	トップクラス * 名詞 サ変名詞 * *
「不明」と判断したもの		
今は解散の	今 * 名詞 普通名詞 * *	今 * 副詞 * * *
国会議員の仕事がなくなり、	仕事 * 名詞 普通名詞 * *	仕事 * 名詞 サ変名詞 * *
直接、政治なり、政治家を	なり * 助詞 接続助詞 * *	なり なる 動詞 * 子音動詞ラ行 基本連用形
モチつきにも人だかり	つき * 名詞 普通名詞 * *	つき つく 動詞 * 子音動詞カ行 基本連用形

ら外している．実験は，読み情報の項目を消してから行なっている．本節での実験では，「タグが誤っている確率」の算出には，2.2節で述べたように用例ベース手法と決定リスト手法を利用する．

まず，1 ～ 5 個の形態素連続における形態素情報を誤りの候補とする[12]．この誤りの各候補に対し，「タグが誤っている確率」と「確信度」と「修正後のタグ」を算出する．(これらの値の算出方法は後述する．) 次に，確信度の大きい誤り候補から順に欲張り法でコーパスを修正する．このとき，各修正箇所には先に算出した「タグが誤っている確率」と「修正後のタグ」を付与しておく．「タグが誤っている確率」が 0.5 より大きい形態素のタグが誤っているものと判定され，「修正後のタグ」に修正される．0.5 以下の形態素のタグは正しいものと判断され，修正の対象とならない．

次に「タグが誤っている確率」と「確信度」と「修正後のタグ」の算出方法を説明する．まず，誤り候補から変更可能な候補をコーパスより取り出す．ここで，変更可能な候補とは表記が同じものである．(先の曖昧形態素の定義と同じ．) 例えば，「ロシア * 名詞 普通名詞 * *」が誤り候補の場合，「ロシア * 名詞 地名 * *」が変更可能な候補として取り出される．ここで，用例ベース手法の場合は，誤り候補のまわりの形態素の状態が最もよく似ている用例を集め[13]，その用例集合で 2.2節で述べた方法で「タグが誤っている確率」と「修正後のタグ」を推定する．「確信度」は「タグが誤っている確率」と「タグが正しい確率」のうち大きい方の値とする．また，決定リスト手法の場合は，以下で説明する 16 個の素性を用いて 2.2節で述べた方法を用いて「タグが誤っている確率」と「確信度」と「修正後のタグ」を推定する．16 個

[12] 本稿では，形態素連続を誤りの候補としているので，枠組みとしては内山の過分割の研究を包含したものとなっている．

[13] 最もよく似ている用例の集め方は，候補の形態素から出発し，それに対して，その前後の形態素の品詞，品詞細分類，残りの全情報を順次追加していき，さらにその隣の形態素からもそのような情報を順次追加する．これを繰り返し，検出される用例が 1 個だけになる直前の状態のときの用例を利用する．

表 7: 形態素誤り修正結果 (上位 20 個，誤り確率 1.0000 ～ 0.9998)

該当箇所	修正前	修正後
けいはんなの経営など、	の * 名詞 形式名詞 * *	の * 助詞 接続助詞 * *
3日夜、札幌発関西	、 * 特殊 記号 * *	、 * 特殊 読点 * *
米朝交渉の駆け引きに悪い	の の * *	の * 助詞 接続助詞 * *
さくらももこの原作で、	の * 連体詞 * * *	の * 助詞 接続助詞 * *
では南方の鉄鉱石の開発	の * 名詞 普通名詞 * *	の * 助詞 接続助詞 * *
中西部、少数民族地区	、 、 * *	、 * 特殊 読点 * *
一年ぐらいの期間をかけて ×	の だ 判定詞 * 判定詞 ダ列特殊連体形	助詞 接続助詞 の * 助詞 接続助詞 * * 接続助詞 助詞
区の同、坂本節子さん	、 * 名詞 サ変名詞 * *	、 * 特殊 読点 * *
年にドルと金との	と * 名詞 普通名詞 * *	と * 助詞 格助詞 * *
をとっても、極めて重要である。	、 、 * *	、 * 特殊 読点 * *
当時としては壮大なスケールの	は * 助詞 格助詞 * *	は * 助詞 副助詞 * *
二日夜、多数のロシア	、 、 * *	、 * 特殊 読点 * *
鈴木いどむが左足で	が * 助詞 接続助詞 * *	が * 助詞 格助詞 * *
にそうだった。	。 。 * *	。 * 特殊 句点 * *
とみたい。	。 。 * *	特殊 句点 。 * 特殊 句点 * * 句点 特殊
などをのどに詰まらせる事故	に だ 形容詞 * ナ形容詞 ダ列基本連用形	に * 助詞 格助詞 * *
のソフトだ。	。 。 * *	。 * 特殊 句点 * *
自転車かごにほうり込んでいく	に にる 動詞 * 母音動詞 基本連用形	に * 助詞 格助詞 * *
については、	、 * 名詞 サ変名詞 * *	、 * 特殊 読点 * *
「ジュラシック・パーク」ぐらいのもの ×	の だ 判定詞 * 判定詞 ダ列特殊連体形	の * 助詞 接続助詞 * * 接続助詞 助詞

の素性については，まず，各形態素の情報として以下の四つのパターンの情報を考え，

1. 情報なし
2. 品詞情報のみ
3. 品詞情報と品詞細分類情報のみ (活用する形態素の場合は，品詞情報と活用形のみを用いる)
4. 形態素情報すべて

この四つのパターン情報を，候補となっている形態素の前後二つの形態素についてあらゆる組合わせを作って，合計 16 個の素性を作り，それを決定リスト用の素性とする．

上記の方法でコーパス修正を行なった．用例ベース手法では 591 個がタグ誤りと検出され，決定リスト手法では 4,054 個がタグ誤りと検出された．その検出されたデータの精度を表 5に示す．また，表中の「ランダム 300 個」は，「誤り確率」のことを考慮せずにコーパスの先頭 300 個を調査したときの精度でほぼ平均精度に相当する．「上位 X 個」は集計したデータを「誤り確率」に基づいてソートし，「誤り確率」の上位のものの精度を調べたものである．「検出精度」は誤り部分を正しく検出した箇所の数を総数で割ったもので，「修正精度」は誤り部分を正しく修正した箇所の数を総数で割ったものである．また表の「不明」は正否がいいがたい場合のものである．検出精度，修正精度の算出では検出，修正を失敗したものとして扱っている．

各タグ誤りの正否は，われわれの直感と経験に基づいて行なっているもので誤っている可能性がある．ここでは，われわれが正しく検出したと判断した場合，正しく修正したと判断した場合，判断が困難で「不明」とした場合などの例を表 6にあげておく．「不明」としたものには，副詞と名詞，サ変名詞と普通名詞，普通名詞と動詞連用形など，タグの定義のゆれに関係しそうなものも含めている．

今回の実験では，表 5のように，用例ベース手法よりも決定リスト手法の方が抽出数，抽出精度ともによかった[14]．決定リスト手法では，抽出総数が約 4,000 で平均精度 (表の「ランダム 300 個」) が 50% 程度あるのでおおよそこの 4,000 のデータを見るだけで 2,000 個の誤りを修正できる計算となる．また，上位での精度は 70% ～ 80% と比較的高く誤りを検出できており，この精度ならば人手でこれをチェックしつつコーパス修正をするのもそれほど負担にならないと思われ，十分実用的にコーパス修正に利用可能ではないかと思われる．

次に，決定リスト手法の上位での修正結果を表 7に示す．該当箇所の欄に×印をつけているものは誤り検出失敗を意味する．検出の上位には，「、 、 * * 」といったコーパス作成中になんらかのデータ作成ミスが生じたのではないかと思われる明らかな誤りも含まれている．「の * 連体詞 * * *」「は * 助詞 格助詞 * *」という

[14] ただし，この結果は本稿での素性の設定状況によるのかもしれず，常に決定リスト手法の方がよいとは限らない．ただし，抽出数は，実はこの修正作業をクローズデータでの解析と見立てた場合，クローズでの不正解の事例数を意味しており，クローズデータでの精度が高くなりやすい用例ベース手法の方がコーパス修正には不向きな手法であるのかもしれない．
また，決定リスト自体クローズでの精度があがりやすい手法である．例えば，素性をすこし増やしてみると，検出数が半減した．素性をどの程度でとどめるかも微妙な問題のようである．

表 8: 読み誤りの修正結果例

該当箇所	修正前	修正後
することが 且 に見えている	目 めに * 名詞 普通名詞 * *	目 め * 名詞 普通名詞 * *
日 朝 国交正常化交渉の	朝 にっちょう * 名詞 地名 * *	朝 ちょう * 名詞 地名 * *
日本の 中小 企業経営者が	中小 なか * 名詞 普通名詞 * *	中小 ちゅうしょう * 名詞 普通名詞 * *
神戸 市の男性	神戸 かんべ * 名詞 地名 * *	神戸 こうべ * 名詞 地名 * *

アノテーターによるミスと思われる誤りもある．「～ぐらいの～」を誤ってコーパス誤りと推定しているが，これはコーパス中の他の誤りが原因となっている．「～ぐらいの～」の「の」はほとんど判定詞「だ」だが，コーパスで格助詞「の」としている箇所が二つあるため，決定リストの一つの素性「～ぐらいの～」における判定詞「だ」のタグがあっている確率が 1 にならず，誤りと検出してしまっている．決定リスト手法の場合は，手法の原理が簡単なために，誤り検出を失敗した場合それならこっちのが誤っているのではないかと推測することが容易なので，誤り検出を失敗したとしても副産物として他の誤りを検出できる可能性が高い．

本稿の実験では読み情報を省いて実験を行なったが，別に読み情報を加えて実験することも可能である．実際に読み情報も含めて実験したときに得られた誤りを表 8に示しておく．中には明らかなアノテータの人為的ミスも見受けられ，本稿でとっている手法が人手による誤りに強い手法のようにも思われる．

4 構文情報の修正

本節では，構文情報の修正結果について述べる．構文情報の修正の試みはすでに乾ら[7] によってなされているが，本稿の決定リストを用いる手法の汎用性を調べる意味でも本稿でも試してみた．構文情報では類似度を定義するのが若干困難であるし，形態素での実験では用例ベースよりも決定リスト手法の方がよかったので，用例ベース手法は試さない．本節の実験では京大コーパスのうち，95 年 1 月 10 日までの約 1 万文のデータを利用する．

以下で修正方法を述べる．ある文節 X の係り先が Y のときに，その文節 X の係り先のタグが正しいかどうかを判定する場合，他の係り先候補を Z_1，Z_2，Z_3 …としたとき，X,Y,Z_i の三つ組みのデータに対し，Y と Z_i の比較で Y が係り先となる確率と Z_i が係り先になる確率を求め (この二つの確率の求め方は後で述べる)，これらの確率の大きい方を「確信度」とし，Z_i が係り先になる確率を「誤っている確率」とし，Z_i を「修正タグ」とする．これをすべての Z_1，Z_2，Z_3 …に対して計算し，このうち，「誤っている確率」が最もおおきい Z_i の「誤っている確率」と「修正タグ」を文節 X に付与する．「誤っている確率」が 0.5 よりも大きい文節の係り先タグは誤っていると判断し，その係り先タグは「修正タグ」に修正する．

次に X,Y,Z_i の三つ組みのデータにおいて，Y が係り先となる確率と Z_i が係り先になる確率の求め方を記述する．この確率の算出には決定リストを利用する．文節情報の A パターンとして以下を定義する．

1. 情報無し
2. 付属語の品詞の情報
3. 付属語の品詞と品詞細分類の情報
4. 付属語の品詞と品詞細分類の情報と，自立語の品詞
5. 付属語の品詞と品詞細分類の情報と，自立語の品詞と分類語彙表の分類番号の上位 5 桁
6. 付属語の品詞と品詞細分類の情報と，自立語の品詞と分類語彙表の分類番号の上位 5 桁と単語自体

文節情報の B パターンとして以下を定義する．

1. 情報無し
2. 自立語の品詞
3. 自立語の品詞と品詞細分類
4. 自立語の品詞と品詞細分類と分類語彙表の分類番号の上位 5 桁
5. 自立語の品詞と品詞細分類と分類語彙表の分類番号の上位 5 桁と単語自体

文節 X には A パターンを文節 Y,Z_i には B パターンを利用し，すべての各パターンの組合わせつまり，6 × 5 × 5 の素性を作る．また，Y と Z_i はどちらが文で先に出現しているかも素性とし，合計 6 × 5 × 5 × 2 の素性をこの決定リストの素性とする．この素性ごとに，コーパスより文節 Y が係り先になる場合の数と，Z_i が係り先になる場合の数を求め，それぞれをその和で割ることでそれぞれの確率値を求める．また，このとき大きい方の確率値を確信度とする．この計算をすべての素性で行なってやり，確信度が最も大きいときの素性の，Y が係り先となる確率と Z_i が係り先になる確率を，X,Y,Z_i の三つ組みのデータにおけるその確率とする．ただし，文節 Y が係り先になる場合の数が 1 でそうでない場合の数が 0 と

表 9: 構文情報の修正結果

	決定リスト手法 (抽出総数 1,456 個)		
	検出精度	修正精度	不明
ランダム 300 個	13%(40/300)	12%(37/300)	9 個
上位 50 個	26% (13/50)	20% (10/50)	2 個
上位 100 個	28%(28/100)	24%(24/100)	2 個
上位 150 個	25%(38/150)	23%(34/150)	2 個
上位 200 個	25%(50/200)	23%(45/200)	5 個
上位 250 個	24%(61/250)	22%(56/250)	5 個
上位 300 個	23%(69/300)	21%(64/300)	6 個

表 10: 正しく構文誤りを修正できたと判断したもの

子供【二人の】家族四人が困ることのないようにとお願いした
【その】建設資金、千二百万カナダドルは、すべてチャリティー活動や寄付で賄い「大半が香港マネーと言ってよい」と古編集局長。
それ以後の名人に勝局への「会心の一手・一局」をピックアップしてもらい、【その】当時の状況を再現してみた。
当たり前のように【存在し、】軽視されてきた水が、いま復権を叫んでいる。
男女間の【賃金や】教育水準の格差などを加味すると、一挙に１７位まで低下することが示されている。

なる素性のデータは削除する．

この方法で実験を行なった結果を表 9に示す．また，正しく構文誤りを修正できたものの例を表 10に示す．表 10の“【”,“】”の記号で囲まれている文節の係り先が，コーパスでは一重下線の文節であったが，二重下線の文節に正しく修正できたことを示している．

表 9のように抽出数がおよそ 1,456で，平均検出精度が 13% なので，この 1,456 のデータから 200 個くらい誤りを検出できると期待される．しかし，修正精度は高いときでも 20% 程度であるので，この手法で構文情報を修正するのは若干忍耐力が必要な状態となっている．それでもその程度の精度で修正ができることとなり，本研究の手法の汎用性を確かめるには十分と思われる．

この構文情報での修正は，乾らの研究でも試みられている．乾らの研究では，１月４日分のデータから 1,241 個箇所を抽出しそのうち 50 個程度誤りを正しく抽出したとなっているので，本研究ではこれよりは高い精度で誤り検出ができていることになる．しかし，乾らの研究では一日分のデータ，つまり，われわれのデータのおよそ 1/9 から 50 個程度の誤りを抽出しているので，再現率は乾らの研究の方が高くなっている．つまり，われわれが採用した決定リストの方法では，適合率が高く再現率が低くなると予想される[15]．コーパス修正はある程度修正してからさらに誤り検出をかけるといった処理をすることも可能なので，上位での精度つまり適合率を重視する方がよいと思われ，決定リストを用いるわれわれの修正手法の方が役に立つのではとも思われる．

5 おわりに

本稿では，まず，内山(6) と乾ら(7) の研究を参考にして，2節でコーパス修正の方法を整理した．その後，その整理した方法を用いて決定リスト手法で形態素情報のコーパス修正を試みた．ソートした結果の上位では 65% ～ 80% という，コーパス修正としてはかなり高そうな精度でコーパス修正が可能であることがわかった．また，4,000 程度の箇所を調べると，その半分，2,000 程度の誤りを抽出できることもわかった．2,000 というのは相当の数でこれを容易に修正できるだけでも有用である．

また，本手法の汎用性を調べるために，構文情報でも決定リスト手法によるコーパス修正を試みた．ソートした結果の上位でも 20% 程度という若干人手修正にはきつそうな値ではあるが，その精度でコーパスの誤り修正が可能であることがわかった．精度は低いがそれなりにコーパスの誤り修正ができており，本手法の汎用性の検証には十分であると思われる．

[15] 決定リストの方法は，素性さえ増やせばクローズでの精度はいくらでも上昇する方法であり，もともとクローズでの精度が高くなりやすい方法である．このため，クローズでの解析誤りに相当する，検出数はもともと少なくなる傾向にあり，誤り検出の再現率も低くなる傾向にある．しかし，この検出数を絞る効果から誤り検出の適合率は高くなるのではと期待している．

参考文献

(1) 森信介, テキストコーパスからの確率的言語モデルの推定, 京都大学工学部博士論文, (1998).

(2) 内元清貴, 関根聡, 井佐原均, 最大エントロピー法に基づくモデルを用いた日本語係り受け解析, 情報処理学会論文誌, Vol. 40, No. 9, (1999).

(3) 村田真樹, 内元清貴, 馬青, 井佐原均, 排反な規則を用いた文節まとめあげ, 情報処理学会論文誌, (2000).

(4) 村田真樹, 内元清貴, 馬青, 井佐原均, 日本語文と英語文における統語構造認識とマジカルナンバー７±２, 言語処理学会誌, Vol. 6, No. 7, (1999).

(5) 内元清貴, 村田真樹, 馬青, 内山将夫, 関根聡, 井佐原均, コーパスからの語順の学習, 自然言語処理研究会 2000-NL-135, (2000).

(6) 内山将夫, 形態素解析結果から過分割を検出する統計的尺度, 言語処理学会誌, Vol. 6, No. 7, (1999).

(7) 乾孝司, 乾健太郎, 統計的部分係り受け解析における係り受け確率の利用法 — コーパス中の構文タグ誤りの検出 —, 自然言語処理研究会 99-NL-134, (1999).

(8) 村田真樹, 内元清貴, 小作浩美, 馬青, 内山将夫, 井佐原均, 位置情報と分野情報を用いた情報検索, 言語処理学会誌, Vol. 7, No. 2, (2000), (to appear).

(9) 黒橋禎夫, 長尾真, 京都大学テキストコーパス・プロジェクト, 言語処理学会第３回年次大会, (1997), pp. 115–118.

(10) 黒橋禎夫, 齋藤由衣子, 坂口昌子, コーパスの作成基準 version 1.6, (1997).

(11) 竹内孔一, 松本裕治, 統計的言語モデルを用いた OCR 誤り修正システムの構築, 情報処理学会論文誌, Vol. 40, No. 6, (1999).

(12) David Yarowsky, Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French, *32th Annual Meeting of the Associtation of the Computational Linguistics*, (1994), pp. 88–95.

(13) 新納浩幸, 複合語からの証拠に重みをつけた決定リストによる同音異義語判別, 情報処理学会論文誌, Vol. 39, No. 12, (1998).

(14) 富浦洋一, 日高達, k-nn 推定法に基づく統語的曖昧さの解消法, 言語理解とコミュニケーション研究会 NLC96-7, (1996), pp. 39–45.

(15) 安藤真一, Yves Lepage, 類似検索機能を備えたツリーバンク構築エディタ, 情報処理学会第 52 回全国大会予稿集, (1996).