自然言語処理　144－15
(2001. 7. 17)

頻度に基づく正の例からの負の例の予測
— 日本語表記の誤り検出と外の関係の文の自動抽出 —

村田 真樹　　井佐原 均

独立行政法人 通信総合研究所
けいはんな情報通信融合研究センター
〒619-0289 京都府相楽郡精華町光台 2-2-2
TEL:0774-95-2424　FAX:0774-95-2429　{murata,isahara}@crl.go.jp

あらまし

正の例からの負の例の予測方法を提案した．この方法では，まず判定すべき事例の一般的な出現確率を正の例から求め，その出現確率が高いにも関わらず正の例のデータに出現しないものを負の例と判定する．また，この手法の具体的応用として日本語表記誤り検出と外の関係の文の自動抽出の研究を行なった．結果は両方の問題ともによく，手法の有効性と汎用性が確かめられた．

キーワード　負の例，表記誤り，スペルチェッカー，外の関係の関係節

Extraction of negative examples based on positive examples
— automatic detection of mis-spelled Japanese expressions and relative clauses that do not have case relations with their heads —

Masaki Murata　　Hitoshi Isahara

Keihanna Human Info-communication Research Center,
Communications Research Laboratory
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
TEL:+81-774-95-2424　FAX:+81-774-95-2429　{murata,isahara}@crl.go.jp

Abstract

We have developed a new method that can be used to extract negative examples based on positive examples. This method calculates the probability of the occurrence of a given example based on positive examples, and then determines an example that has a high probability of occurrence but does not appear in the set of positive examples to be a negative example. We have applied this method to two important problems: automatic detection of mis-spelled Japanese expressions and automatic extraction of relative clauses that do not have case relations with their heads. The results were positive for both problems, thus confirming the effectiveness and the applicability of our method.

key words　Negative Examples, Spelling Errors, Spell Checker, Non-Case Relational Relative Clause

1 はじめに

本稿では正の例からの負の例の予測の問題を扱う．例えば，日本語文の表記誤りの検出を考える場合，大規模な既存のコーパスをすべて正しいと仮定すると，その既存のコーパスを正しい文，つまり，正の例と考え，この情報を用いて，日本語の表記誤り，つまり，負の例を予測し抽出することになる．この意味で正の例からの負の例の予測の問題は，実際の日本語文の表記誤りの検出に役に立つ重要な課題となる．

しかし，正の例のみからの学習は一般に難しいことでしられている(1)．また，正の例だけでなく，負の例 (例えば文の表記誤りの検出の問題の場合は実際の表記誤りの例) がある場合は，それらを教師信号としてサポートベクトルマシン法などの教師あり機械学習手法で扱うことができ(2)，高精度な処理が期待できるが，正の例のみのデータでは，教師あり機械学習手法を用いることができない．本稿ではこのような難しい正の例のみからの学習の問題を正の例のデータでの頻度情報などを用いることにより扱うものである．

本稿での正の例からの負の例の推測の基本的アイデアは以下のとおりである．

- まず，正の例か負の例か判定すべき未知の事例の一般的な出現確率を，なんらかの方法で算出する．
- 次に，この出現確率で既知の正の例のデータに出現しないことが不自然である場合，つまり，一般的な出現確率が高く当然正の例のデータに出現するであろう状態にも関わらず既知の正の例のデータに出現しない場合，負の例の度合いが高いと推測する．

本稿では実際に上記の手法を，日本語表記誤り検出の問題と，外の関係の文の抽出の問題に適用し，手法の有効性を確かめる．外の関係の文とは，連体修飾節の動詞と被修飾要素の名詞とが格関係にない文のことをいう．例えば，「負の例をみつけることは難しい」のような文は，「みつける」と「こと」の間にガ格やヲ格などの格関係がないために，外の関係の文とされる．ここで格関係にある連体修飾節を正の例とするとき，外の関係の文は負の例となる．格関係にある動詞と名詞はコーパス中に多く存在するためこの情報を正の例として，負の例の外の関係の文を予測すると外の関係の文の自動獲得が行なえるというわけである．

本稿の主張点をあらかじめ整理しておくと以下のようになる．

- 本稿では正の例からの負の例の予測を，正の例での頻度情報を用いることで実現する．
- 正の例からの負の例の予測の本稿の手法を，実際に日本語表記誤り検出の問題と外の関係の文の自動獲得の問題に適用し手法の有効性および汎用性を確認した．

2 正の例からの負の例の予測方法

正の例からの負の例の予測方法としては，単純なものだと，既知の正の例のデータに現れなかったものを，すべて負の例とすることが考えられる．しかし，実際には未出現の正の例の存在が考えられるために，この方法を用いると，多くの正の例を負の例であると判定してしまうことになる．

そこで，本稿では正の例での頻度情報を用いることで，もう少し精度よく負の例を予測することを試みる．

本稿での正の例からの負の例の推測方法は以下のとおりである．

1. まず，正の例のデータ D と，正の例か負の例か判定すべき未知の事例 x が与えられる．
   未知の事例 x が正の例のデータ D に含まれるときは正の例であると判定する．そうでないときは，次の処理を行なう．
2. 判定すべき事例 x の一般的な出現確率 $p(x)$ を，なんらかの方法で計算する．
   (この方法の具体例は後述する．)
3. この出現確率 $p(x)$ を使って判定すべき事例 x が正の例のデータ D に出現する確率 $P(x)$ を求める．このとき，正の例のデータが n 個でありそれぞれが独立であることを仮定すると，1 回試行して事例 x が出現しない確率は $1-p(x)$ であり，これが n 回連続して起こるということから，事例 x が正の例のデータ D に出現しない確率は $(1-p(x))^n$ となり，事例 x が正の例のデータ D に出現する確率 $P(x)$ は $1-(1-p(x))^n$ となる．
   ところで，ここでの「事例 x が正の例のデータ D に出現しない」というのは，コーパスが小さいために確率的に出現しないということが保証されたことを意味するため，「事例 x は正の例でありうる」という意味になる．逆に，「事例 x が正の例のデータ D に出現する」というのは，確率的にはコーパスに当然出現すべきということになり，それなのに実際は出現しなかったということで矛盾が生じることになり，一般的な出現確率 $p(x)$ か種々の独立の仮定が否定されることになる．ここで，「事例 x が正の例である場合は，一般的な出現確率 $p(x)$ および種々の独立の仮定が正しい」と新たに仮定すると，この矛盾により「事例 x は正の例でありえない」が導出されることになる．つまり，「事例 x が正の例のデータ D に出現する確率 $P(x)$」は，「事例 x が正の例でありえない確率 $P(x)$」を意味することになる．そういう意味で，$P(x)$ は負の例の度合いを意味するものとなる．以降，この $P(x)$ を**負の例の度合い**と呼ぶことにする．
4. 事例 x の負の例の度合い $P(x)$ が大きいほど，事例 x の負の例の度合いが大きいと考える．

上記の処理では 2 の「判定すべき事例 x の一般的な出現確率 $p(x)$ を，なんらかの方法で計算する．」の部分

だけ具体的な説明をしていない．この一般的な出現確率 $p(x)$ の算出方法の一例を以下に示す．

- ここでは，特に正の例のデータは二項関係 (a,b) からなるものと仮定する．このとき，この二項の a, b が互いに独立であると仮定すると，二項関係 (a,b) の出現する確率 $p(x)(=p(a,b))$ は，a, b の正の例のデータでの出現確率を $p(a), p(b)$ とするとき[1]，その積の $p(a)\times p(b)$ になる．

つまり，各データを二項関係とし各項を独立と仮定することで，各データの一般的な出現確率を，その部分の確率により計算するのである．

本稿での正の例からの負の例の予測方法は以上のとおりである．以降，実際にこの手法を具体的に利用した研究として，3節に日本語表記誤り検出の研究を，4節に外の関係の文の抽出の研究を述べる．

3 日本語表記誤り検出システム

本節では，正の例からの負の例の予測技術の適用例として，日本語表記の誤り検出をとりあげる．

3.1 背景と先行研究

単語の表記誤りに限った話では，日本語での誤り検出は，英語の場合に比べてはるかに難しいものである．英語の場合は，単語でわかち書きされているために，基本的に単語辞書と，単語末の変形の規則を用意しておくことで，ほぼ高精度に単語のスペルチェックを行なうことができる(3)．これに対して，日本語の場合は，単語でわかち書きされていないために，単語の表記誤りに限ったとしても扱うのが困難である．

また，表記の誤りとしては単語表記の誤りの他に，助詞の「て」「に」「を」「は」の運用誤りなどの文法的な誤りも存在する．

日本語の表記誤りの検出の先行研究としては主に以下のものがある．

- 単語辞書やひらがな連続を登録した辞書や，連接の条件を記述した辞書に基づいて，表記誤りを検出する(4, 5, 6)．単語辞書やひらがな連続を登録した辞書にないものがあらわれると表記誤りと判定したり，連接の条件を記述した辞書において満足されない連接の出現が存在すると表記誤りと判定する．
- 文字単位の ngram を利用した確率モデルに基づいて，各文字列の生起確率を求め，生起確率の低い文字列が出現する箇所を表記誤りと判定する(7, 8, 9)．

上記にあげた研究のうち，後者の ngram 確率を利用する手法は，主に OCR 誤り訂正システムにおける，表記誤り検出に用いられているものである．OCR 誤り訂正システムの場合は，前提として表記誤りの出現率が 5 ～ 10% と高く，普通に人がものを書くときに誤る確率より高く，表記誤りの検出の再現率，適合率は高くなりやすく，比較的容易な問題設定となる．

[1] 厳密には第一項の a の出現確率 $p(a)$ は，正の例のデータの第一項に a が出現した個数を総数で割ったものである．$p(b)$ の計算も同様に正の例のデータの第二項のデータを用いて計算する．

	負	の	事	零	の	検	出
trigram に		-1	-1	-1			
よる得点			-1	-1	-1		
Total		-1	-2	-2	-1		

図 1: 竹内らの手法の模式図

また，この後者の研究の中で最も良さそうに思われる竹内らの方法(9) を，以下で詳しく説明する．

- 表記誤りを検出したいテキストを頭から，一文字ずつずらしながら，3 文字連続を抽出し，それのコーパス (正しい日本語文の集合) での出現確率が Tp 以下の場合，その各 3 文字連続に -1 を加えていく．与えられた値が Ts 以上となった文字を誤りと判定する．ここで，$Tp=0, Ts=-2$ とする．

ここで，$Tp=0$ としているために確率はわざわざ求める必要はなく，コーパスにその 3 文字連続が出現するかいなかを調べるということをするだけでよい．$Tp>0$ とした場合はコーパスに出現するものがあっても誤りと判定するものとなる．しかし，出現確率が低くとも出現していればそれは誤りとしなくてよいだろうから，$Tp>0$ はよくなく竹内らの方法の $Tp=0$ の設定はよいものと思われる．

ここで図 1を用いて例を用いて竹内らの方法を補足説明する．「負の事零の検出」という日本語表現に対して誤り検出を行なうことを考える．このとき，頭から「負の事」「の事零」といった 3 文字連続を切り出し，これらがコーパスにあるかどうかを調べ，それがなければその 3 文字に -1 を与える．この場合「の事零」「事零の」がなかったため，図のように得点が与えられ，結果として -2 点となった「事」と「零」の部分が誤りと判定される．この竹内法はコーパスに高頻度に出現する文字 3gram をうまく組み合わせて誤りを検出する方法となっている．

しかし，結局のところ，行なっていることはコーパスにその表現が存在するかいなかであり，コーパスでの確率や頻度を用いないものとなり，辞書にないものがあらわれると誤りとする前者の研究群とよく似たものとなっている．

3.2 本手法

次にわれわれの表記誤り検出方法を説明する．

本稿でのわれわれの表記誤り検出方法は，2節で述べた正の例からの負の例の予測方法を用いて行なう．

正の例としてはコーパス (正しい日本語文の集合) を用いる．二項関係としては任意の連続する 1 ～ 5gram の二つの文字列とする．基本的な考え方は，この二つの文字列の連接チェックをコーパスで行なうことにより誤りを

表 2: 本提案手法による誤り検出結果 (上位 10 個)

負の度合い	前方文脈	二項関係 x_{max}		後方文脈
0.99999	説明した方	法で	を	用いることができる
0.99999	した」などの表現が用	いら	る	。
0.99999	咲く」を含む文でのは	次	よう	な対応関係を
0.99999	主題と述語が与え	ら	た場	合に意味ネットワーク
0.99999	して出力する方が適切	が	であると	考え、
0.99999	文として出力する方が	適切	が	であると考え、
0.99999	例えば、	図	よう	なのネットワークから
0.99999	語が与えらた場合に意	味	ネ	ットワーク
0.99999	合にはすべての並列節	とい	て出	力した
0.99998	合にはすべての並列節	と	いて出	力した

表 1: 誤りを含む例文

1	自然な つながりがもつようにする 必要がある。
2	「~する」「~した」などの表現が用 いらる。
3	例えば、図 ような のネットワークから、
4	「咲く」を含む文 でのは次ような 対応関係を
5	可能な連体節 がである 場合は、この連体節を
6	説明した方法 でを 用いることができる
7	文として出力する方が適切 がであると 考え、
8	主題と述語が与 えらた 場合に意味ネットワーク
9	可能な場合にはすべての並列節 といて 出力した

検出するのである．その二つの文字列が連接できる場合を正，できない場合を負と考える．連接できる場合の正の例はコーパス (正しい日本語文の集合) から取ることができるため，正の例からの学習の問題となり，2節で述べたわれわれの手法が適用できる．

われわれの手法の詳細な手順は以下のとおりである．

1. 文の頭から，文字の各すき間ひとつひとつを，連接チェックの対象とし，以下の処理を行なう．
   (a) 対象としている文字のすき間に前接する 1 ~ 5gram の文字列 a と，後接する 1 ~ 5gram の文字列 b を取り出し，この任意のペア $x = (a, b)$ (25 個できる) を作る．
   ここで，連接 ab がコーパスにあるかどうかを調べ，ある場合は，そのペアは除いて以下の処理を行なう．また，すべてのペアがコーパスにあった場合は，連接するものと判断し，連接は妥当なものと判断して処理を次のすき間に移す．
   (b) 各ペア x ごとに 2節の負の例の度合い $P(x)$ を求める．
   (c) もっとも，$P(x)$ の値が高いときのその値を，P_{max}，また，x を x_{max} とする．
   (d) $P(x_{max})$ の値が大きいすき間ほど，妥当でない連接の可能性が高いと判定し，処理を次のすき間に移す．

上記の手順は，2節で述べた手法を若干拡張したものになっている．二項関係を各箇所 25 種類を作成し，それぞれで負の例の度合い $P(x)$ を求め，これのもっとも大きいときの値を最終的な判定に用いているものになっている．つまり，連接チェックのパターンとして 25 種類を用意し，この中でもっとも負の例の度合いの大きいパターンを最終チェックに利用するということになっている．妥当性のチェックという場合は，各種のチェック機構を用い，そのうち一つでもひっかかると妥当でないと判断するのがよい．本手法は，それに近く，多くのチェックパターンを用意し，その中で一番チェックにひっかかったところのチェックの評価を最終評価に用いるということになっている．

ところで，コーパスとしては，実際に誤り検出をかけるデータ自身も用いることができる．この場合，自分自身を用いると当然自分自身のデータによりチェックの対象となる表現は必ず 1 回以上検出されることになる．このため，出現頻度は 1 をひいて用いることになる．(この手法は leave one out 法と等価である．) しかし，この方法では，全コーパスを通じて二回以上まったく同じ誤りが出現する場合，その誤りは検出できないという問題があるので，利用には注意が必要である．しかし，本稿の日本語表記誤り検出の実験では，すべてにおいてこの自分自身のデータを用いる方法で行なった．(比較手法として用いる竹内らの方法でもこの方法で実験した．)

3.3 実験

本手法の有効性を確かめるために実験を行なった．

まず，白木らの研究[6] であげられていた表 1に示す 9 つの誤り例を検出できるかどうかの実験を行なった．誤り部分に下線をひいている．実験で正の例としたデータは，毎日新聞[10] の 91 年から 98 年のデータである．

われわれの手法の結果の上位 10 個を表 2にあげる．上位では負の度合いが極めて高くほとんど上限の 1 に近

表 3: 1 文字削除データでの誤り検出精度

	再現率	適合率
本手法		
上位 50 個	2.88%	92.00%
上位 100 個	5.31%	85.00%
上位 200 個	9.69%	77.50%
上位 300 個	12.88%	68.67%
上位 500 個	18.56%	59.40%
上位 800 個	24.06%	48.12%
上位 1,200 個	29.12%	38.83%
上位 1,600 個	32.38%	32.38%
上位 3,000 個	39.75%	21.20%
上位 5,000 個	47.38%	15.16%
上位 10,000 個	57.38%	9.18%
上位 20,000 個	67.81%	5.42%
上位 50,000 個	81.38%	2.60%
上位 100,000 個	89.31%	1.43%
竹内らの手法		
総検出数 5,295 個	25.31%	7.65%

表 4: 1 文字置換データでの誤り検出精度

	再現率	適合率
本手法		
上位 50 個	3.12%	100.00%
上位 100 個	5.94%	95.00%
上位 200 個	11.56%	92.50%
上位 300 個	16.62%	88.67%
上位 500 個	25.25%	80.80%
上位 800 個	34.25%	68.50%
上位 1,200 個	42.44%	56.58%
上位 1,600 個	48.06%	48.06%
上位 3,000 個	61.38%	32.73%
上位 5,000 個	70.81%	22.66%
上位 10,000 個	81.12%	12.98%
上位 20,000 個	87.62%	7.01%
上位 50,000 個	94.44%	3.02%
上位 100,000 個	97.81%	1.57%
竹内らの手法		
総検出数 5,944 個	60.94%	16.40%

いことがわかる．また，「意味ネットワーク」以外は誤り検出に成功していることがわかる．1 の例を除いたすべての事例を上位 25 個以内に検出できていた．1 の例は，「自然な つながりがもつようにする 必要がある。」となっていて確かにおかしい文であるが上位では検出できなかった．コーパスのあらゆるひらがな連続を辞書に登録しそれに無いひらがな連続を誤りとする白木らの方法では，8, 9 の例ができないとなっていたが，本手法ではそれらも上位で検出できている．

また，比較として竹内らの方法でも実験を行なった．竹内らの方法では，11 箇所を誤り候補として検出した．正しく検出できたものは 3 例であり再現率に問題があるといった感じであった．

次に作為的に誤り箇所を生成したデータを用いた擬似的な実験を行なった．この実験は京大コーパス[11] にある毎日新聞の 95 年の 1 月 17 日までの 16 日間の約 2 万文 (892,655 文字) で行なった．実験は 1 文字削除，1 文字置換，1 文字挿入の三種類を行なった．この三つの実験は独立に行なった．また，各実験では各日に 100 個の誤りをランダムな箇所に生成し，それぞれ合計 1,600 個の誤りを作成した．このとき，誤り箇所の前後 10 文字以内に他の誤りが出現しないような条件を設けた．また，置換，挿入時に新たに置かれる文字は，京大コーパスの 91 年から 94 年のデータでの文字の出現頻度分布に比例する条件でランダムに決定した．作成した誤りが 1,600 文字で元の文字数が 892,655 文字であるから，誤り文字の出現率は 0.18% で，558 文字に 1 つの割合で誤りが生じていることになる．また，実験で正の例の学習データとしたものは，毎日新聞の 91 年から 94 年のデータである．また，実験は 1 日分のデータを一つの記事 (データ) として入力した．つまり，3.2節で説明した自分自身のデータも含めて行なうという方法の自分自身のデータは，この 1 日分となる．

この実験はわれわれの手法だけでなく，竹内らの方法でも行なった．これらの結果を表 3から表 5に示す．ここでは再現率と適合率を評価に用いた．再現率は正解の数を誤りの総数 1,600 で割ったものを意味し，適合率は正解の数を検出数で割ったものを意味する．表の「上位 X 個」は負の度合いでソートしたデータの上位 X 個までの精度を意味する．また，正解の判定は表記誤りをしている 1 文字をきっかり指摘しなくて一文字前後にずれていても正しく検出したと判定する．また，すでに正解不正解の判定をした事例の一文字前後の事例はその事例の指摘が正解でない場合は以降の判定から除いている．

結果の表から以下のことがわかる．

- 当然のことだが，上位 X 個の X が増えるにつれて，つまり，検出数が増えるにつれて，再現率が上昇する．
- 上位 1,600 個のところを見ると，再現率と適合率が一致する．これは誤りの総数と検出数が一致するためである．

この地点で調べると，おおよそ，1 文字削除データで精度が 1/3 で，1 文字置換 / 挿入データで精度が 1/2 であることがわかる．

これは，先に述べたように今回の擬似的データでは

表 5: 1 文字挿入データでの誤り検出精度

		再現率	適合率
本手法			
上位	50 個	3.12%	100.00%
上位	100 個	6.00%	96.00%
上位	200 個	11.62%	93.00%
上位	300 個	16.69%	89.00%
上位	500 個	24.88%	79.60%
上位	800 個	33.88%	67.75%
上位	1,200 個	41.94%	55.92%
上位	1,600 個	47.19%	47.19%
上位	3,000 個	60.44%	32.23%
上位	5,000 個	69.88%	22.36%
上位	10,000 個	80.62%	12.90%
上位	20,000 個	88.69%	7.09%
上位	50,000 個	95.25%	3.05%
上位	100,000 個	98.12%	1.57%
竹内らの手法			
総検出数	5,944 個	62.12%	16.72%

558 文字に 1 つの割合で誤りが生じている状態であるので，おおよそ 400 字詰原稿用紙 1 枚半に一つ誤りがあるというときには，1 文字削除が約 1/3 の確率で約 1/3 を検出でき，1 文字置換あるいは挿入がおおよそ半分の確率で半分を検出できることを意味する．

しかし，一般に誤りの出現率が減ると，誤りでないのに誤りと指摘する誤りが生じ，精度は低下する．誤りの出現は正しいものの出現に比べ大幅に小さいので，一般的には単純に，誤りの出現率が半分になると，誤った検出になる原因部分が倍になり精度は半分になると考えるとよい．

- 次に竹内らの方法と比較する．

  竹内らの方法では，検出の際にソートする基準となる尺度がないため，上位だけを調べるなどといったことができない．これに対し我々の方法では上位の精度よく検出されたところだけを手早く直す，といったことができる．

  また，竹内らの方法では再現率が固定で，1 文字削除で 25%，他のもので，60% であり，多くの誤りを必ず見過ごすものとなっている．

  また，基本的な精度も検出数のよくにた上位 5,000 くらいで比較すると，われわれの手法の方が高いものとなっている．

以上のことから，本提案手法でそれなりに誤り表記の検出が行なえることがわかった．また検出性能はそれほど高くはないが，負の度合いでソートできるために，簡単に修正できそうなものだけ手早く修正することができるという利点も存在するので，検出性能の低さを理解した上でならば実際に利用することは現段階でも十分考えられる．また，本稿では対象を日本語にしたが，本稿の手法は英語などの他の言語における文法エラーチェックなどにも用いることができるだろう．

ところで本稿での提案は負の例の検出手法であったが，上記の結果より，これは，表記誤り検出の課題で十分その力を発揮したといえるだろう[2]．

4 外の関係の文の抽出

本節では，正の例からの負の例の予測技術の適用例として，外の関係の文の抽出を行なう．

4.1 背景と先行研究

外の関係の文とは，例えば以下のような埋め込み文の節の動詞とその係り先の名詞の間に格関係が成立しないものをいう(12)．

「負の事例を抽出することは難しい．」

例えば，上の文の場合，「負の事例を抽出すること」という関係節では，「抽出する」という動詞とその係り先の「こと」という名詞の間で，「ことが抽出する」や「ことを抽出する」などのような格関係が成立しない．このため，このような文は外の関係と呼ばれる．これとは逆に格関係が成立する文は内の関係と呼ばれる．

外の関係の文は上記のような形式的なものの他に，

「さんまを焼くけむり．」

などといった複雑な構造をしたものもある．

外の関係の文を抽出する研究としては，阿部川らのもの(13) や Baldwin のもの(14) や表のもの(15) がある．阿部川らの方法は，連体修飾関係と格関係で，それを構成する動詞の異なり数の分布に大きな違いがあることに着目し，その分布の違いを KL- 距離を用いて評価することで外の関係の文を特定する．Baldwin は，連体節に関して緻密な研究を数多く行なっており，外の関係になりやすい名詞をあらかじめ抽出するなどしてその情報を利用した人手ルールに基づく方法を用いた研究から，格フレーム情報を含む広範な情報を属性とした教師あり機械学習を用いて外の関係を特定する研究などを行なっている[3]．また，表は埋め込み文の日英翻訳のために，格フレームの情報を用いて外の関係か内の関係かを判定する研究(15) を行なっている．

外の関係の文の抽出の研究とは直接関係はないが，「りんごを」と「食べる」が共起し得るかなど，二つの語が格関係として共起可能であるかどうかを，単語の共起情報，つまり，正の例の情報のみを利用して，判別分

[2] 本稿の本論とは関係のないことだが，毎日新聞のデータには「をを」という誤りが多く存在しているようである．また，「ッッ」などのだぶり誤りも多そうである．このような表記誤りは列挙もしくは規則化して対処するのがよいだろう．また，本稿の手法はこの種の代表的な誤りの検出にも役に立つ．

[3] さらには外の関係自体もいくつかの分類に意味的に分割し，それらの分類を機械学習により推定するといった研究も行なっている．

表 6: 本手法の外の関係の検出精度

	再現率	適合率	正解率
上位 10 個	3.75%	100.00%	70.46%
上位 20 個	7.12%	95.00%	71.38%
上位 30 個	8.99%	80.00%	71.38%
上位 40 個	10.86%	72.50%	71.38%
上位 50 個	13.48%	72.00%	71.84%
上位 100 個	20.97%	56.00%	70.69%
上位 150 個	28.46%	50.67%	69.54%
上位 200 個	33.33%	44.50%	66.78%
上位 300 個	39.70%	35.33%	59.20%
総検出数 393 個	42.32%	28.75%	50.11%

表 7: 教師あり学習手法に基づく外の関係の検出精度

	再現率	適合率	正解率
上位 10 個	3.75%	100.00%	70.46%
上位 20 個	7.49%	100.00%	71.61%
上位 30 個	11.24%	100.00%	72.76%
上位 40 個	14.98%	100.00%	73.91%
上位 50 個	18.73%	100.00%	75.06%
上位 100 個	36.33%	97.00%	80.11%
上位 150 個	53.18%	94.67%	84.71%
上位 200 個	66.67%	89.00%	87.24%
上位 300 個	77.53%	69.00%	82.41%
総検出数 870 個	100.00%	30.69%	30.69%

析の手法により求めようとしている興味深い研究[16]もある．ただし，この研究では入力の情報はすべて 1 か * の形で扱われており，本稿で提案する手法のような正の例のデータ集合での頻度情報などは用いられていない (* は未定の情報を意味する)．また，手法自体もまったく異なる．

4.2 本手法

次にわれわれの外の関係の文の抽出方法を説明する．

本稿でのわれわれの外の関係の文の抽出方法は，2節で述べた正の例からの負の例の予測方法を用いて行なう．正の例としてはコーパス (正しい日本語文の集合) から knp[17] などを用いて取り出した格関係にあるとされる名詞と動詞の対のデータを用いる．また，二項関係は名詞と動詞の対を考える．基本的な考え方としては，高頻度に出現する名詞と動詞の対にも関わらず，上記のあらかじめ取り出した正の例に存在しなければそれらは外の関係であろうと判定する．

われわれの手法の詳細な手順は以下のとおりである．

1. コーパスから knp などを用いて大量の連体節の動詞とそのかかり先の名詞の組 x=(a,b) を取り出す．これらのデータが外の関係かどうか判定されるものとなる．
   また，コーパスから knp などを用いて大量の格関係にあるとされる名詞と動詞の組 y を取り出す．y は正の例のデータとして用いられる．
2. 判定する連体節の動詞とそのかかり先の名詞の組 x=(a,b) に対して，次のことを行なう．
   x が y の集合に含まれる場合，x は正の例と判定され外の関係でなく内の関係であると判断する．
   x が y の集合に含まれない場合は，x が名詞と動詞の二項関係からなるものと考えて 2節の負の例の度合い P(x) を求め，この値が大きいほど，負の例の度合いが大きいと判定し，外の関係である可能性は高いと判定する．

4.3 実験

提案手法の有効性の確認のために実験を行なった．この実験はわれわれの格解析の研究で用いている少量のデータ (1,530 個) のうちの連体節にかかわるデータ (870 事例) を用いて行なった．このデータでは各事例が外の関係であるかいなかがふられているために，自動で精度を求めることができる．外の関係の事例はこのうち 267 個であった．正の例のデータとしては毎日新聞の 95 年を除く 91 年から 98 年までの 7 年分のデータを用いた．

また，この実験は 870 事例に外の関係であるかの情報がふってあるため，教師あり機械学習手法でも実験ができる．そこで比較のために機械学習手法を用いたもので行なった．機械学習での実験はこの 870 事例を 10 分割し，そのクロスバリデーションにより精度を求めた．

これらの結果を表 6と表 7にあげる．評価は再現率と適合率と正解率で行なった．再現率は正しく外の関係を特定できた数を外の関係の総数 267 で割ったものを意味し，適合率は正しく外の関係を特定できた数を検出数で割ったものを意味する．正解率は，その正解率を求める地点までの事例を外の関係と判断した場合の全事例 870 個での外の関係と内の関係の区別の正解精度である．表の「上位 X 個」は負の度合いでソートしたデータの上位 X 個までの精度を意味する．また，教師あり機械学習手法では文献[18] の素性集合で機械学習手法に最大エントロピー法を用いた[4]．また，出力された事例は最大エントロピー法が出力する解に対する確率値によりソートした．表 6の本手法では総検出数 393 と元の事例総数 870 より少なくなっているが，これは正の例に同じ名詞・動詞対があるか，名詞，動詞のどちらかが正の例のデータに一回も出現せず正の例であると判断したものを除いた結果を示しているためである．

実験結果から以下のことがわかった．

[4] 文献[18] のように，精度は最大エントロピー法よりもサポートベクトルマシン法の方がよかった．しかし，ここでは事例をソートするために，確率値を出力する最大エントロピー法の方の結果を示す．

- 機械学習手法を用いた結果の方が精度が圧倒的によい.

  しかし，本手法は，上位10個まででは精度は100%であり，負の事例の教師信号がなく正の事例だけからでもそれなりに外の関係を抽出できることがわかる.
- また，機械学習手法の場合は学習データにあった事例と良く似た外の関係の文しか抽出できない可能性があり，本手法ではそのようなものも抽出できる可能性があるため，精度が低いからといってすぐに役に立たないと判断されるわけではない.
- また，正の例に同じ名詞・動詞対があるために，外の関係と判定できなかったものに以下のものがある.

  「目に見える形で」

  この例では「形」と「見える」が格関係になく外の関係の文となっているが，実際のコーパスでは「形に見える」など意味的には異なるが格関係として構成できる場合があるため，正の例と判定してしまった.

  そのような誤りが多かったため，精度が低くなっているものと思われる.

以上のように本手法の精度は教師あり学習手法よりは精度の低いものであった．しかし，教師あり学習の場合は，正の例だけで推定する手法に比べて，負の例の情報もあるために高い精度を得やすいと思われる．また，本手法は全般的に精度が低いとはいえ，上位での適合率は高い．外の関係の出現率は30.7%であり本手法では上位10個で連続して正解しているが30.7%の確率のものを10回連続生じる確率は0.0000074であり，これは偶然生じるようなことではない．このことから，本手法がランダムな選択よりは効果があることは確実である．また，既知の負の例の情報を用いない手法としては，それなりによい結果だとも考えることができる.

5 おわりに

本稿では正の例からの負の例を予測する手法を提案した．実際にこの方法を日本語表記誤り検出と外の関係の文の抽出に用いた．両方の問題ともに，負の例の度合いでソートした結果の上位では高い適合率で負の例を検出できることを確認した．二つの問題で成果が出たということで本手法の汎用性も確かめられた．つまり，他の多くの正の例からの負の例を予測する問題も同様に解くことができることが予想される.

しかし，外の関係の実験では負の例も与えた教師あり学習手法の精度の方が圧倒的によかった．このことから，できることならば正の例だけでなく負の例も用意して教師あり機械学習手法を用いるのがよいということがわかる．このため，安易に本手法の適用を考えるのではなく，本当に正の例，負の例の教師信号を用意することができないかをよく確かめた上で，本手法を利用する必要がある.

ところで，本手法の正の例からの学習は表記誤り検出に用いたように，文の適切性の判定に利用できる．われわれの言い換えの先行研究[19]では，コーパスに1回以上出現することを条件とするような手法をいくつか用いていたが，そのかわりに本稿の手法を利用して本稿の負の度合いが小さいことを条件とするといった改善を行なうことができる．このようにすることで，コーパスに1回も出現しないような正しい表現も妥当と判定できるようになる．といっても，生成の適合率を高めるためには1回出現は条件として用いた方がいいとも考えられるので，状況に応じた手法の選択が必要となる.

参考文献

(1) 横森貴, 篠原武, 形式言語の学習 — 正の例からの学習を中心に —, 情報処理学会誌, Vol. 32, No. 3, (1991), pp. 226–235.

(2) 村田真樹, 馬青, 内元清貴, 井佐原均, サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳, 電子情報通信学会 言語理解とコミュニケーション研究会 NLC2000-78, (2001).

(3) Karen Kukich, Techniques for automatically correcting words in text, *ACM Computing Surveys*, Vol. 24, No. 4, (1992), pp. 377–439.

(4) 納富一宏, 日本語文書校正支援ツール hsp の開発, 情報処理学会研究発表会 (デジタル・ドキュメント), (1997), pp. 9–16.

(5) 川原一真, 山本幹雄, コーパスから抽出された辞書を用いた表記誤り検出法, 情報処理学会 第54回 全国大会, (1997), pp. 2-21–2–22.

(6) 白木伸征, 黒橋禎夫, 長尾真, 大量の平仮名列登録による日本語スペルチェッカの作成, 言語処理学会 年次大会, (1997), pp. 445–448.

(7) 荒木哲郎, 池原悟, 塚原信幸, 2重マルコフモデルによる日本語文の誤り検出並びに訂正法, 情報処理学会自然言語処理研究会 NL97-5, (1997), pp. 29–35.

(8) 松山高明, 渥美清隆, 増山繁, n-gram による ocr 誤り検出の能力検討のための適合率と再現率の推定に関する実験と考察, 言語処理学会 年次大会, (1996), pp. 129–132.

(9) 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).

(10) 毎日新聞社, 毎日新聞 1991-1998, (1998).

(11) 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.

(12) 寺村秀夫, 連体修飾のシンタクスと意味 — その1～その4, 日本語・日本文化, Vol.4-7, 1975-1978, ().

(13) 阿部川武, 白井清昭, 田中穂積, 徳永健伸, 統計情報を利用した日本語連体修飾節の解析, 言語処理学会 年次大会, (2001), pp. 269–272.

(14) Timothy Baldwin, Making lexical sense of japanese-english machine translation: A disambiguation extravaganza, Technical report, (Tokyo Institute of Technology, 2001), Technical Report, ISSN 0918-2802.

(15) 表克次, 埋め込み文の日英翻訳方式, 鳥取大学卒業論文, (2001).

(16) 富浦洋一, 田中省作, 日高達, 不完全データに対する判別分析と語の共起性推定への応用, (2001).

(17) 黒橋禎夫, 日本語構文解析システム KNP 使用説明書 version 2.0b6, (京都大学大学院情報学研究科, 1998).

(18) 村田真樹, 井佐原均, 機械学習を用いた日本語格解析 — 教師信号借用型と非 借用型, 情報処理学会 自然言語処理研究会 2001-NL-144, (2001), (to appear).

(19) 村田真樹, 井佐原均, 言い換えの統一的モデル — 尺度に基づく変形の利用 —, 言語処理学会第7回年次大会ワークショップ論文集, (2001).