決定リストとアダブーストを利用した訓練データ中の誤り検出*

○ 新納浩幸

茨城大学 工学部 システム工学科

1　はじめに

自然言語処理の個々の問題を分類問題として定式化し，帰納学習の手法を利用して解決するアプローチは，大きな成功をおさめている．そこでは学習で使われる訓練データが重要な位置付けにある．なぜなら帰納学習の手法には大規模かつ信頼性の高い訓練データが必要とされるからである．しかし訓練データは通常人手で作成されるために，その構築に多大なコストがかかるという問題もある．そのような背景から用意できた訓練データの優劣が，学習する規則の優劣を決していることも多い．

本論文では訓練データ中の誤りを検出する手法を提案する．前述したように訓練データは人手で作成されるため，誤りが混在しやすい．本手法により訓練データの質を高め，結果として学習で得られた規則の精度向上が行える．ここでは，日本語単語分割を題材にする．この場合，訓練データは単語分割されたコーパスに相当する．つまり本手法は，単語分割されたコーパスから単語分割誤りを検出する手法とも見なせる．

ここでは訓練データ中の誤り検出の手がかりとして2つの点に注目する．1点目は，決定リスト [2] により単語分割規則を学習した場合，決定リストの個々の証拠の持つ正解率はリストの上位にある証拠ほど高いことを利用する．この性質を利用すると，学習させた決定リストにより訓練データを単語分割した場合，リストの上位にある証拠によって判断された単語分割はほぼ正解になるはずである．そのために，逆に不正解になった場合には，訓練データの方が誤っている可能性が高い．2点目はアダブーストの利用である．アダブーストは弱学習器の組み合わせから強学習器を作る手法である．アダブーストでは，まず，訓練データから学習器（この場合，単語分割規則）を学習する．次に，学習した規則により訓練データを解析し，誤った事例に重みをつけて訓練データを再構築し，再び学習器を学習する．この操作をn回繰り返し，n個の学習器を作る．最終的な判別はn個の学習器の重み付き多数決により行う．上記手順中，重みを付けていってもなかなか学習できないデータ，つまり非常に大きな重みを持つ事例は誤りである可能性がある [1]．この性質を利用する．

実験では京大コーパスを訓練データとして本手法を試した．誤りの可能性の高い上位100個所を出力し，それらを確認したところ，その多くは単語分割の判定が微妙な個所であった．また完全に人間のケアレスミスである21個所の誤りを検出できた．

*Detection of errors in training data by using the decision list and Adaboost

2　決定リストによる日本語単語分割

本手法は訓練データから単語分割のための決定リストを学習し，その決定リストを使って訓練データを解析することで誤り個所を検出する．本章では決定リストによる日本語単語分割方法を概説する．

日本語単語分割は入力文 ($s = c_1c_2\cdots c_n$) の各文字の間 (c_i と c_{i+1} の間の地点 b_i) に単語境界を置く（クラス +1）か置かない（クラス −1）かの分類問題としてとらえることができる．

分類問題は帰納学習の手法を利用して解決できる．帰納学習には様々な手法があるが，どの手法が優れているかは問題に依存するので，一概には言えない．ここでは決定リスト [2] を利用する．

2.1　決定リストの構築

決定リストの分類規則は証拠とクラスの組の順序付きの表である．ここで証拠とは属性とその属性の値の組である．実際の判別はリストの上位のものから順に，その証拠があるかどうかを調べ，その証拠があれば，それに対応するクラスを出力する．

決定リストの作成は概ね以下の手順による．

step 1 属性を設定する．

例えば n 個の属性を $att_1, att_2, \cdots, att_n$ とする。

step 2 訓練データから証拠とクラスの組の頻度を調べる．

訓練データ中のある事例の属性 att の値が a であるとし，その事例のクラスが C だとする．その場合，(att, a) という証拠とクラス C の組 $((att, a), C)$ の頻度に 1 を足す．これを訓練データ中の全事例に対する全属性について行う．

step 3 証拠の判別力と分類クラスを導く．

$((att, a), C)$ の頻度が f_C であった場合，f_C の最大値を与える $\hat{C}$ が証拠 (att, a) に対する分類クラ

スとなる．またそのときの判別力 $pw(att, a)$ は以下で定義される．

$$pw((att, a)) = \log \frac{f_{\hat{C}}}{\sum_{C \neq \hat{C}} f_C}$$

step 4 判別力の順に並べる．

全ての証拠と分類クラスの組を判別力の大きい順に並べる．これによって作成できた表が決定リストである．

2.2 属性の設定

各文字間 b_i がどのクラスに属するかを判断する材料が属性である．本論文では b_i の属性として，表 1 の 7 種類を用意した．

表 1: 設定した属性

属性	値	
att_1	文字列	$c_{i-1}c_ic_{i+1}$
att_2	文字列	$c_ic_{i+1}c_{i+2}$
att_3	文字列	$c_{i-1}c_i$
att_4	文字列	c_ic_{i+1}
att_5	文字列	$c_{i+1}c_{i+2}$
att_6	字種の接続関係 1	((c_iの大分類字種), (c_{i+1}の大分類字種))
att_7	字種の接続関係 2	((c_iの細分類字種), (c_{i+1}の細分類字種))

6，7 番目の属性として，字種の情報を利用している．ここでは字種を大分類と細分類の二つの観点から分類した．字種の大分類は 6 番目の属性，字種の細分類は 7 番目の属性で利用した．

字種の大分類は平仮名，カタカナ，漢数字，漢字，英数字，アルファベット，○，◯ の計 9 種類である．字種の細分類は大分類の平仮名の部分をその文字自身にしたものである．

3 単語分割の誤り検出

3.1 決定リストを利用した誤り検出

決定リストの個々の証拠の持つ正解率はリストの上位にある証拠ほど高い．この性質を利用すると，決定リストにより訓練データを単語分割した場合，リストの上位にある証拠によって判断された単語分割はほぼ正解になるはずである．そのために，逆に不正解になった場合には，訓練データの方が誤っている可能性が高い．この性質を利用して誤りの可能性の高い事例を選出する．

本論文の決定リストは各文字間に単語分割境界を置くか置かないかを判断する．訓練データ中の文章中の各文字間にはその正解が付与されている．そのため決定リストにより訓練データを単語分割した場合，どの文字間の判断が誤り，しかもそれは決定リスト中のどのランクの証拠により判断されたかがわかる．この判断が誤った部分で，しかも上記のランクが高いものが，訓練データ中の誤りがある可能性が高い事例である．

3.2 アダブーストを利用した誤り検出

弱学習器の組合わせから強学習器を作る手法をブースティングという．アダブーストはブースティング手法の代表的手法であり，その有効性は多くの研究で示されている．

アダブーストのアルゴリズムを図 1 に示す [1]．アダブーストでは，まず，訓練データ中の各事例に等しい重みを置く．この訓練データから学習器を構築し，この学習器を利用してその訓練データの各事例の判別を行う．判別の誤った事例に重みを課して，再び，学習器を構築する．これを n 回繰り返し，n 個の学習器を作る．最終的な判別は n 個の学習器の重み付き多数決により決める．

Given: $(x_1, y_1), \cdots, (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{1, -1\}$
Initialize $Di(i) = 1/m$
For $t = 1, \cdots, T$

- Train weak learner using distribution D_t
- Get weak hypothesis $h_t : X \rightarrow Y$ with error

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i]$$

- Choose $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$
- Update:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if \quad h_t(x_i) = y_i \\ e^{\alpha_t} & if \quad h_t(x_i) \neq y_i \end{cases}$$

where Z_t is a normalization factor

Output the final hypothesis:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

図 1: アダブースト

ここでは訓練データから学習される学習器は決定リストである．重みは頻度として与える．決定リストを作成する際には訓練データ中の各事例には重みがついているとして，その重みが決定リスト作成の step 2 で各証拠と正解の組に付加する数値とする．図 1 のアルゴリズムでは正規化するために重みの総和が 1 になっているが，ここでは重みの最小値が 1 となるようにして計算を簡単にした．このため最初の決定リストを作成する際の各事例の重みは 1 であり，2 回目では正解の事例の重みは 1 で変化せず，不正解の事例の重みが大きくなる．

アダブーストは最も分類困難な事例に重みを集中させるので，重みの大きな事例は例外的な事例であることが多い．この性質を利用して訓練データ中の誤りを検出する．本論文の場合，アダブーストにより最適な個数の決定リストを作成し，その重み付き多数決より訓練データ中の事例の判別を行う．ここで誤った判別を起こした事例は，次のステップで非常に大きな重みをつけられる．結果的に，最適な個数の決定リストからの重み付き多数決による判定が誤ったものが，訓練データの誤りの可能性が高いものである．

3.3 決定リストとアダブーストの組み合わせによる誤り検出

決定リストとアダブーストを利用して訓練データ中の誤りの可能性のあるものを検出する方法をそれぞれ述べた．ここでは，それらを組み合わせる方法を述べる．

まず決定リストを利用して訓練データ中の各事例にどのランクの証拠により判別が行われたかを記録する（結果1）．ここではその判別が正しいか，誤りかには注目しない．次に前述したアダブーストを利用した訓練データ中の誤り検出方法を利用して，誤りの可能性のある事例を列挙する（結果2）．最後に（結果2）の各事例に対して（結果1）に記載されているランクを付与する．このランクの高い事例を順に並べたリストが，誤りの可能性の高い順に並べた事例のリストとなる．本論文では上位100種類を出力する．

4 実験

本手法を京大コーパスに対して適用した．京大コーパスは人手で構文構造のタグがつけされたコーパスであり，単語分割の情報もつけられている．

京大コーパスでは38,383文に対して単語分割のタグがつけられている[1]．単語分割境界が存在するかしないかの判定を行う個所から訓練データの事例を構築できる．結果，1,635,505個の事例を構築できた．これが訓練データとなる．

ここから本論文で述べた決定リストを作成し，訓練データの各事例がどのランクの証拠で判定されたかを記録した．次に，アダブーストを適用した．その結果を図2に示す．

縦軸は正解率，横軸はブーストティングの繰り返しの回数を表す．3つの決定リストを作成し，それらの重み付き多数決による判別の場合が，最もよい精度を出した．そのときの誤り数は3,349個であった．これらの個所がアダブーストにより検出した誤りの可能性のある事例である．次にこの3,349個の事例の先に付与しておいたランクを調べ，そのランクの高いもの100個を出力した．これが本手法の出力である．

[1]ここではコーパス中の記号EOSの数を文の数としている．句点“。”の数ではないことを注意しておく．

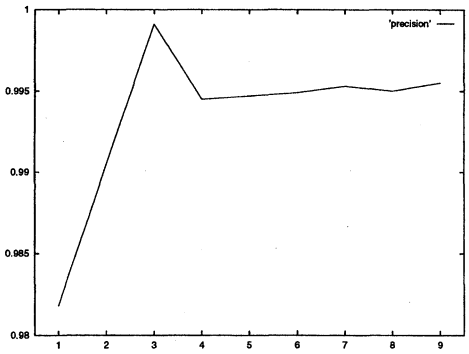

図2: ブースティングによる正解率の変化

この100個を手作業で分類した結果を以下に示す．

(1) 完全な誤り (21個) これは訓練データの方が明らかに誤っているものである．以下に例を示す．下線の部分が検出できた誤りである．

- 立候補/する/人/が/いるか/と/思えば/
- お/金/の/無駄遣/い のこ/と/を/
- が/あった/」/と知事/へ/の/
- ＪＲ/西日/本/も/
- クリーチャー/と/の/間にあった/愛/

(2) 連語表現 (25個) 複数の単語が組み合わさって1つの単語を構成している連語は，訓練データでは一単語として記述されている．しかし，本手法では連語に対して単語が分割されると判断して誤りを指摘しているものが多い．以下に例を示す．

- 政治家/ばかり/が/大手を振って/きた
- この/日記/も/気をつけ/ない/と
- つまり/下手をすると/計画/倒れ
- 二言目には/「/補助金/行政/など
- はんらん/は/目に見えて/いる

(3) カタカナ表現 (38個) 複数の英単語に対応するカタカナ表現は訓練データでは分割するケースと分割しないケースで揺れがある．そのために本手法と分割の判断が異なり，誤りとして指摘しているものが多い．以下に例を示す．

- キングファハド/スタジアム (本手法では1単語)
- ノート/メーカー (本手法では1単語)

- ビールメーカー (本手法では “ビール/メーカー”)
- プロ/フィギュアスケート (本手法では 1 単語)
- シーズン/オフ (本手法では 1 単語)

(4) その他 (16 個) これは誤りの指摘が完全に間違いであるものである．全て平仮名表現であった．

分類 (1) の部分を正解（有効な検出）と見なすと，正解率は 0.21 である．手作業で作られたものからの誤り検出であることを考えれば，有益な値である．また分類 (2), (3) も正しい分割の判断が微妙であるために，それらも考慮すると，本手法が指摘する誤り個所の多く（ 84% ）は有益であった．

5 考察

誤り検出システムの評価は，単純な正解率（検出した誤りが本当に誤りである割合）では測れない．誤り検出システムはその目的から考えて，そのシステムを利用することで，手作業で誤りを見つけるよりも効率的に誤りを見つけることが出来ればよいのであり，その効率性が誤り検出システムの評価になるはずである．本システムでは，上位 100 個の検出で 21 個が完全な誤りであった．訓練データ中の判断すべき分割個所が約 163 万個であることを考えれば，手作業で誤りを発見することは不可能であり，その点で本手法は有用であると考える．

また訓練データ中の誤ったタグを検出できるシステムが，高精度なタグ付けシステムであるとは限らないことを注意しておく．理由は 2 つある．1 つは，誤り検出システムが誤っていないと判断した部分の判定を，誤り検出システム単独で出せるかどうかの保証がないからである．通常，訓練データにタグをつける基盤のシステムが存在し，その基盤システムの出力の誤り部分だけを誤り検出システムで修正するというアプローチをもったシステムが，高精度なタグ付けシステムとなる可能性がある．2 つ目は先に述べた正解率の問題である．誤り検出システムでは正解率だけが重要であるわけではない．そのために，基盤システムと誤り検出システムを組み合わせても精度の向上が起こるとは限らない．

ここで提案した誤り検出システムも単純に基盤システム（ここでは JUMAN）の精度をあげることは出来ない．正解率が低いからである．100 個のうち 21 個の修正が正しくても，副作用として 79 個の誤りを新たに発生させるので，結果的に精度が下がる．ただし，本手法で，訓練データが正しいのに誤りと検出した部分は，カタカナ表記，平仮名表記となる場合がほとんどである．この点を利用して，これらの部分では誤りとして検出しないことにすれば正解率は若干は上がると思われる．

ここで作成した誤り検出システムが，単独で高精度な単語分割システムとして利用できるかどうかは，学習した決定リストの精度にかかっている．この部分では幾つかの実験を行っているが [4]，JUMAN 以上の正解率は得られていない．本質的に本手法では単語分割の判定を，問題の地点の前後数文字で判断しているために，明らかに精度の限界がある．辞書の情報を利用できるようにするのが鍵である．

また本論文では単語分割タグに対する誤り検出を行ったが，他のタグにも応用できる．その際の注意すべき点としては，分類問題の解決を通してそのタグがつけられることと，その分類問題に決定リストが利用できることである．例えば，同音異義語問題 [5] などはここでの手法が直接利用可能である．一般に新聞記事には同音異義語の誤りは含まれないという仮定から，同音異義語問題の訓練データとしては直接新聞記事が使われるが，その仮定は保証されてはいない．誤りを修正した訓練データにより精度が向上することも示されている [3]．

6 おわりに

本論文では，訓練データ中のタグの誤りを検出することを目的に，2 つの手がかりを利用する手法を提案した．1 つ目の手がかりは適用した決定リストの証拠の判別力である．訓練データ中の事例を，判別力が高い証拠によって判別したにも関わらず，その判定が誤りであるとすれば，その事例のクラスは誤りである可能性が高い．2 つ目の手がかりはアダブースの重みである．アダブースの学習器を作成する繰り返しの中で多大な重みが課せられてゆく事例は誤りである可能性が高い．ここでは日本語単語分割問題を対象にした．実験では京大コーパスを対して本手法を適用した．誤りの可能性の高い順に上位 100 個を出力した結果，21 個が実際の誤りであり，本手法の有用性が示された．

参考文献

[1] Yoav Freund, Robert Schapire (訳: 安倍直樹). ブースティング入門. 人工知能学会誌, Vol. 14, No. 5, pp. 771–780, 1999.

[2] David Yarowsky. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *32th Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, 1994.

[3] 伊吹潤, 西野文人. 誤りを含み得るコーパスからの校正支援用データの整備. 言語処理学会第 5 回年次大会, pp. 177–180, 1998.

[4] 新納浩幸. 決定リストを弱学習器としたアダブーストによる日本語単語分割. 自然言語処理, Vol. 8, No. 2, 2001 (to appear).

[5] 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, pp. 3200–3206, 1998.