自然言語処理 120－14
(1997. 7. 25)

誤り駆動型学習とシソーラスを用いた文書自動分類

山崎 毅文
NTT コミュニケーション科学研究所
〒239 横須賀市 光の丘 1-1
yamazaki@cslab.kecl.ntt.co.jp

イド ダガン
Bar Ilan University
dagan@cs.biu.ac.il

本稿では、誤り駆動型学習アルゴリズム WINNOW を用いたテキスト自動分類手法について述べる。WINNOW は、事例中にノイズを含む場合や分類に無関係な属性が多数存在する場合に対して、効率的に働くことが知られており、上記特徴を持つテキスト分類の問題に対して、有効に働くことが期待できる。本提案手法では、テキストを表す特徴として、単語だけでなくシソーラスによって付与される意味カテゴリーも合わせて利用する。シソーラス利用によって生じる 2 つの問題点、特徴空間の次元増加による過適用の問題及び多義の問題を解決する手段として、Filtering 手法及び関連度に基づく多義性解消手法を提案する。RWCP テキストコーパスを用いた分類実験により、提案手法の妥当性を示す。

Mistake-driven learning with thesaurus for text categorization

Takefumi Yamazaki
NTT Comunication Science Laboratories
1-1 Hikariooka Yokosuka Kanagawa 239 Japan
yamazaki@cslab.kecl.ntt.co.jp

Ido Dagan
Bar Ilan University, Israel
dagan@cs.biu.ac.il

This paper extends the mistake-driven learner WINNOW, which has been highly studied in the theoretical machine learning literature, to better utilize thesauri for text categorization. In our method not only words but also semantic categories given by the thesaurus are used as features in a classifier. New filtering and disambiguation methods are used as pre-processing to solve the problems caused by the use of the thesaurus. In the experiment we test RWCP corpus and verify our method.

1 はじめに

大量の電子化されたテキストの蓄積／流通に伴ない、これらの大量のテキストの中から有用な情報をいかに効率的に抽出できるかが重要な課題となっている。この課題に対処する一つの手段は、テキストをその意味内容に基づいて予め決められたカテゴリーに分類することである。

この分類作業の自動化を実現するのが「テキスト自動分類」であり、予めカテゴリー分類されたテキストを利用して適当な分類器を作成し、その分類器に基づいて、新たに入力されたテキストのカテゴリー分類を行なう。この分類器をいかに精度良いものにするかという課題に対し、従来から様々な方式が提案されているが[Lew92, LG94, ADW94]、分類精度に関してまだ改良の余地を残している。

本稿では、最近学習研究の分野において、研究が進展している誤り駆動型学習アルゴリズム WINNOW[Lit88] を用いて分類器を生成する手法について述べる。WINNOW は、2層ニューラルネットワークの重みを計算する乗算型更新アルゴリズムであり、最近の研究で理論的な解析が進んでおり、その振舞いがよく理解されている [Lit88, Lit95]。

WINNOW は、事例中にノイズを含む場合や分類に無関係な属性が多数存在する場合に対して、効率的に働くことが知られており、上記特徴を持つテキスト分類の問題に対して、有効に働くことが期待できる。

また本稿で提案する手法は、テキストを表す特徴として、単語だけでなくシソーラスによって付与される意味カテゴリーも合わせて利用する。シソーラス利用によって生じる新たな問題、即ち、特徴空間の次元増加による過適用の問題と多義性の問題とを解決する手段として、Filtering 手法と関連度に基づく多義性解消手法を提案する。

本稿の構成は、以下の通りである。まず、WINNOW の学習対象モデルである線形分類モデルについて述べ、今回利用した Balanced WINNOW について述べる。次に、シソーラス付与及びその利用において生じる問題点について述べ、その解決手段としての Filtering 手法及び多義性解消手法について述べる。次に、RWCP テキストコーパスを用いた分類実験について述べ、最後にまとめと今後の課題について述べる。

2 誤り駆動型学習 WINNOW によるテキスト分類

2.1 線形分類モデル

一つのテキスト d は、通常、特徴 f_i の集合として、$d = \{f_1, f_2, ...f_m\}$ のように表現される。ここで、m はテキスト中に現れる特徴の数を表す。一般的に、特徴はテキスト中の単語 w で表現される。テキスト d 中に現れる特徴 f の強さを s(f,d) で表す。ここでは、強さ s(f,d) を d 中に現れる f の出現回数とする。

テキストをどの該当分類カテゴリーに振り分けるか判断するため、d に対する各分類カテゴリー c 毎のスコア Fc(d) を用いる。このスコアがある閾値を越えた場合に、テキストに対し該当分類カテゴリーを付与する。

線形分類モデルでは、各特徴の各分類カテゴリー c に対する重みを、$w_c = (w(f_1,c), w(f_2,c), ..., w(f_n,c)) \equiv (w_1, w_2, ..., w_n)$ として定義する。ここで、n は特徴数の総数、w(f,c) はカテゴリー c に対する特徴 f の重みを表す。以上の定義から、Fc(d) は強さと重みの積：$Fc(d) = \sum_{f \in d} s(f,d) \cdot w(f,c)$ で表現できる。

最終的に、線形分類モデルにおける学習は、新しいテキストを分類するのに最適な重みベクトルを得ることである。

2.2 誤り駆動型学習アルゴリズム: WINNOW

誤り駆動型学習アルゴリズム WINNOW [Lit88] は、前述の線形分類モデルにおける重みを学習する方法である。理論的解析により、無関係属性が多く存在する場合でも有効に働くこと、また、その誤り率の上限は、分類に関連ある属性の数に対し線形、全属性数に対して log の割合で増加することが知られている。実世界への応用も、最近活発である [Blu95][GR96]。

本章において、まず、WINNOW アルゴリズムの基本的な動きを Basic WINNOW として述べ、次にその拡張版である Balanced WINNOW について述べる。

2.2.1 Basic WINNOW

WINNOW アルゴリズムは、訓練事例が一つ入力される毎に、分類判定を行ない、その分類結果が誤まった場合のみ、重み更新という形でフィードバックを行なうオンライン型の学習アルゴリズムである。

WINNOW によって、各特徴の重みベクトルを得る手順を以下に述べる。まず、学習対象分類カテゴリー c を定め、訓練事例を予め付与されている分類カテゴリーに基づき、正事例と負事例とに分ける。

分類モデルは、分類カテゴリー c 毎に N 次元の重みベクトル $w_c = (w_{c_1}, w_{c_2}, ..., w_{c_n})$ を保持する。まず最初に、ベクトルの初期化が行なわれ、全ての特徴に対し、ある正の重みを与える。本アルゴリズムは、閾値 θ と 2 つの更新パラメータ（増進パラメータ $\alpha > 1$ 及び、降下パラメータ $0 < \beta < 1$）の計 3 つのパラメータを持つ。

特徴ベクトル $(s_1, ..., s_m)$ を持つ入力事例に対し、本アルゴリズムは、$\sum_{j=1}^{m} w_{c_j} \cdot s_j > \theta_c$ 時のみ、1（正事例）と判定する。ここで、w_{c_j} は、事例中に出現する特徴である。本アルゴリズムは、判定を誤った場合のみ次の 2 つの戦略で重みを更新する。(1) アルゴリズムが 0（負事例）と判定し、分類ラベルが 1 の場合、事例中に出現する特徴の重みを α 倍することにより増進する（$w \leftarrow \alpha \cdot w$）。(2) アルゴリズムが 1 と判定し、分類ラベルが 0 の場合、事例中に出現する特徴の重みを β 倍

することにより減少させる（$w \leftarrow \beta \cdot w$）。この重み更新の操作を全訓練事例が収束するまで、あるいは、決められた回数だけ繰り返す。

2.2.2 Balanced WINNOW

線形分類モデルでは、重みの合計が閾値が越えるか否かで、クラス判定しているため、前述のBasic WINNOWでは、長さの長いテキストは、正解が負事例の場合でも、単にその長さ即ち特徴数の多さ故に、閾値を越えるという問題点があった。

Balanced WINNOWでは、各特徴は、w^+, w^-の2つの値を持ち、最終的に各特徴の重みは、この2つの差で表現されるので、負の値が取れる。特徴ベクトル$(s_1, ..., s_m)$を持つ入力事例に対して、次の式が成立する場合のみ、1（正事例）と判定する。

$$\sum_{j=1}^{m}(w_{c_j}^+ - w_{c_j}^-) \cdot s_j > \theta_c$$

アルゴリズムは特徴の重みを次の戦略で更新する。

(1) 正解が正事例に対して負事例と誤った場合、重みの正の部分のみ増進し（$w^+ \leftarrow \alpha \cdot w^+$）、負の部分を減少させ（$w^- \leftarrow \beta \cdot w^-$）、結果として式1における$s_j$の係数を増加させる。(2) 正解が負事例に対して正事例と誤った場合、重みの正の部分を減少し（$w^+ \leftarrow \beta \cdot w^+$）、負の部分を増加させる（$w^- \leftarrow \alpha \cdot w^-$）。

本戦略によって、上記式において特徴の重みは負の値を取ることが可能であり、本提案手法では、線形分類モデルの学習アルゴリズムとしてBalanced WINNOWを採用した。

3 シソーラスを利用した特徴空間生成

前述の通り、線形分類モデルで用いられる特徴空間は、通常単語で構成される。特徴空間を拡張する一つの手段として、句表現[Lew92]や単語の2-gram、3-gram[CS96]の利用、また、単語クラスタリングによる手法[李97]が考えられる。しかし、これらの手法は、特徴をあくまでテキスト自身から生成し、テキストに現れない情報は利用しないという点で、限界がある。提案手法では、特徴空間を単語及びシソーラスによって得られるその単語の意味カテゴリーから構成する。シソーラスの利用で単語より一般的な項目で特徴空間が構成できるため、データスパース問題の解決が期待でき、また他の言語知識源の利用により未知単語への対処が可能である。

実験では、次の2種類のシソーラスを用いた。一つは、対象分野を特に限定しないで作成された一般シソーラスである「分類語彙表」[国立64]であり、もう一つは、機械翻訳における意味解析用に作成された「ALTシソーラス」[池原93]である。

両シソーラスは、収録語数、その粒度の点で異なり、「分類語彙表」は、32,600単語を収録しており、約800種類の意味カテゴリーが付与されており、シソーラスの深さは約4段である。一方、「ALTシソーラス」の方は、約200,000単語が収録されており、約2,000種類の意味カテゴリーが付与され、シソーラスの深さは12段である。また、両シソーラス共、各々の単語が持つ意味カテゴリーの数は、2つ以上の場合もあり、この場合、多義による曖昧性の増加に繋がる。

4 Filtering手法

分類モデルの学習において、分類カテゴリーの特定に無関係な特徴を予め除去できれば、分類モデルの精度向上が期待できる。特に、シソーラス利用による特徴数の増加に伴い、分類モデルが訓練事例に対して過適用(Overfitting)する可能性が増え、これが分類予測の精度悪化に繋がる。

この過適用問題を解決するため、無関係な特徴を除去する、2種類のFiltering手法を考案した。一つは、WINNOWの学習結果得られる重みの値に基づく手法であり、もう一つは、χ^2値に基づく手法である。

WINNOWの学習結果として得られる重みの値に基づくFilteringは、以下の操作によって行なわれる。分類カテゴリー毎に、訓練事例が収束するまであるいは決められた回数分、WINNOWにより各特徴の重みを学習する。次に、得られた重みの値に応じて、関係特徴、無関係特徴を区別し、関係特徴のみ用いて、再度WINNOWによる学習を行なう。

一方、χ^2値に基づくFilteringは、以下の操作によって行なわれる。まず、各分類カテゴリー毎に、各特徴に対して、以下の4つの値（該当特徴が出現するテキスト集合の中で、該当分類カテゴリーを持つテキスト数：N_{r+}、持たないテキストの数：N_{n+}、また該当特徴が出現しないテキスト集合の中で、該当分類カテゴリーを持つテキスト数：N_{r-}、持たないテキストの数：N_{n-}）を求める。以上4つの値から、下の式に従って、各特徴のχ^2値を計算する。

$$\chi^2 = \frac{N(N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+} + N_{r-})(N_{n+} + N_{n-})(N_{r+} + N_{n+})(N_{r-} + N_{n-})}$$

このスコアの値が大きい程、該当分類カテゴリーの特定に際する有用度が高いと判断できる。そこで、この値がある決められた閾値以下の特徴を無関係特徴として除去し、関係特徴のみ用いて、WINNOWによる学習を行なう。

5 曖昧性解消による意味カテゴリーの選択的利用

先に述べたように、シソーラス利用によって得られる各単語の意味カテゴリーの数は一つとは限らない。分類モデルに与える特徴として、特定の意味カテゴリーに絞

らずに、曖昧性があるまま全ての意味カテゴリーを利用すると、訓練事例中のノイズが増え、分類精度を悪化させる可能性がある。

例えば、単語「路線」は2つの分類カテゴリー『政治』『交通』の両方で頻出する多義語であり、その意味カテゴリーは、ALT シソーラス上で「AC 交通路」「AC 線」「AC 形勢」の3つである[1]。その中で「AC 交通路」は『交通』に関係あるが、『政治』には無関係であり、また、「AC 形勢」はその逆である。よって、多義を絞らず単純に全ての意味カテゴリーを利用すると、学習モデルに無関係な特徴を与えることになってしまう。

[河合92] では、文の統語解析結果を用いて、多義性解消を行なっているが、我々は、テキストの分類カテゴリー情報と事前処理で得られる特徴と該当分類カテゴリーの関連度を利用して、多義性解消を行なう。

意味カテゴリーは、その対象テキストが属する分類カテゴリーに関連性のある他の単語からも生成される。例えば、「AC 形勢」は分類カテゴリー『政治』で頻出する単語「動向」「非常事態」からも生成される意味カテゴリーであり、一方、「AC 交通路」は分類カテゴリー『交通』で頻出する単語「新幹線」「東海道」からも生成される。よって、『政治』『交通』の分類カテゴリーを持つテキストから、「AC 形勢」は『政治』と強い関連を持ち、また「AC 交通路」は『交通』と強い関連を持つことが解るはずである。この関連度を利用すれば、対象テキスト中に単語「路線」が現れた時、そのテキストの分類カテゴリーが『政治』であれば、意味カテゴリーとして「AC 形勢」を選択し、『交通』であれば、「AC 交通路」を選択して用いることが可能である。

特徴と分類カテゴリーの関連度として、Filtering 手法と同様に、WINNOW による学習結果で得られる重み、及び χ^2 値 を利用する。

2つの関連度計算の方法に従って、「路線」「AC 交通路」「AC 形勢」「AC 線」の、『政治』『交通』に対する関連度を計算した結果を、表1及び表2に示す。表1はWINNOW による学習結果で得られる重み であり、表2は χ^2 値である。これらの値に基づいて、候補カテゴリーの中でスコアが最も大きいものを最適なカテゴリーとして選択する。両表から解るように、いずれの手法を利用しても、『政治』に関連する意味カテゴリーとして「AC 形勢」が、また、『交通』に関連する意味カテゴリーとして「AC 交通路」が選択でき、多義性解消が可能であることが解る。

[1]単語と区別するため、ALT シソーラス上の意味カテゴリーを"AC"で表現する。

特徴属性	“『政治』”	“『交通』”
路線	3.8732	1.8557
AC 交通路	1.8557	3.6650
AC 形勢	2.7460	-0.3928
AC 線	-1.5979	0.0000

表 1: WINNOW 学習による重みを利用した特徴と分類カテゴリーの関連度

特徴属性	“『政治』”	“『交通』”
路線	81.4497	24.1962
AC 交通路	20.4147	58.1479
AC 形勢	72.5792	0.1949
AC 線	3.3743	5.8368

表 2: χ^2 値による特徴と分類カテゴリーの関連度

6 実験結果

6.1 実験設定

実験には、RWCP テキストコーパス[豊浦96] を用いた。本コーパスは、１９９４年度の毎日新聞データ約３万件の記事に、国際十進分類法のＵＤＣコード[情報94] を付与したものである。その３万記事の中から、まず、予め決めた１３種類の分類カテゴリー[2]を持つ記事を選択し、さらにそれ以外のカテゴリーを持つ記事をランダムに選択し、両者をマージすることで、約９０００テキストから成るサンプルデータを生成し、その中から、訓練事例及びテスト事例として各々、３０００テキストずつ選択した。

対象テキストは、NTT の形態素解析システム ALT-JAWS によって、形態素解析をし、名詞、固有名詞を抜きだし、その後、2種類のシソーラスによる意味カテゴリー付与を行なった。WINNOW に与えるパラメータは、$\alpha = 1.05$、$\beta = 1/\alpha$、$\theta = 10$ に設定し、利用する特徴は頻度による簡単な Filtering を行なった。

6.2 評価方法

分類精度の評価尺度として、的中率、網羅率、及びF値[vR79] を利用した。各分類カテゴリー毎に、分類モデルと正解の正事例及び負事例の数から、下記の数をカウ

[2]選択した分類カテゴリーは、1. 刑法 (343) 2. 国際関係 (327) 3. 軍事技術 (623) 4. 精密機械 (681) 5. スポーツ (796) 6. 交通 (656) 7. 演劇 (792) 8. 野球 (796.357) 9. 作物 (633) 10. サッカー (796.332) 11. TV放送局 (654.197) 12. 政府 (328) 13. 内閣 (328.13) である。() は、対応する UDC コードを表す。

ントする。

- a: 正解が正事例かつ分類モデルが正事例と判断した数
- b: 正解が負事例かつ分類モデルが正事例と判断した数
- c: 正解が正事例かつ分類モデルが負事例と判断した数

的中率、網羅率は、次のように定義される：

- 的中率$(P) = a/(a+b)$
- 網羅率$(R) = a/(a+c)$

また、F値は、下記のように的中率、網羅率の重み付き結合で定義される。

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

実験では、β=1の場合のF_β（即ちF_1= 2a/(2a+b+c)）を利用した。

6.3 結果

6.3.1 シソーラス及びFilteringの効果

表3に3つの実験設定、(1)単語のみ(2)単語 + ALTシソーラス(3)単語 +分類語彙表について、Filteringをしない場合、χ^2に基づくFiltering、WINNOWの重みの値に基づくFiltering(Weight-Filtering)を利用した場合の分類精度を示す。また、表4に、Filteringをしない場合、F値に関して、どちらの実験設定が分類精度が良いかを分類カテゴリー毎に比較した結果を示す。

(1) vs. (2) Which wins?	(1) vs. (3) Which wins?	(2) vs. (3) Which wins?
3 < 10	3 < 8	9 > 3

表4: 分類カテゴリー毎の比較

表3において、Filteringを利用する／しない、いずれの場合も、シソーラスを利用した(2)(3)の場合の方が、単語のみの場合(1)に比べて、分類精度が良く、シソーラス付与の効果が確認できる。また、χ^2-Filtering、Weight-Filtering、いずれのFiltering手法を用いても、3つの実験設定全てにおいて、分類精度が向上することが解る。

一方、表4の結果から、シソーラス利用の方が単語のみの場合に比べて、全てのカテゴリーで精度が良いわけではなく、いずれのシソーラス利用の場合も1 3カテゴリー中、3カテゴリー（3. 軍事技術 10. サッカー 12. 政府）については、精度が下回っていることが解る。

Filteringを利用した場合も、この傾向は変わらなかった。よって、Filteringは、シソーラス利用により生じる多義性の問題等の解決に特に効果があるのでなく、分類精度の絶対値を押し上げる効果があるとみなせる。また、(2)と(3)を比較した場合は、ALTシソーラスを利用した場合(2)の方が良いことが解る。シソーラスの細かさ、粒度、収録語数の違い等が、精度の差の原因と思われる。

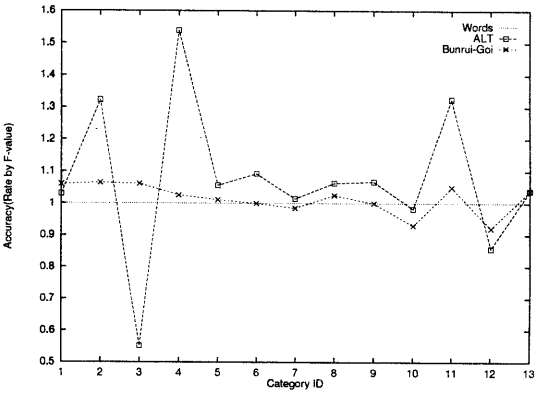

図1: 分類カテゴリー毎の相対精度(Filtering無し)

図1に、Filteringしない場合における、分類カテゴリー毎のF値に関する(1)に対する(2)(3)の相対比を示す。この図から、3つのカテゴリー（2. 国際関係 4. 精密機械 11. TV放送局）において、特にカテゴリー利用の効果が大きいことが解る。

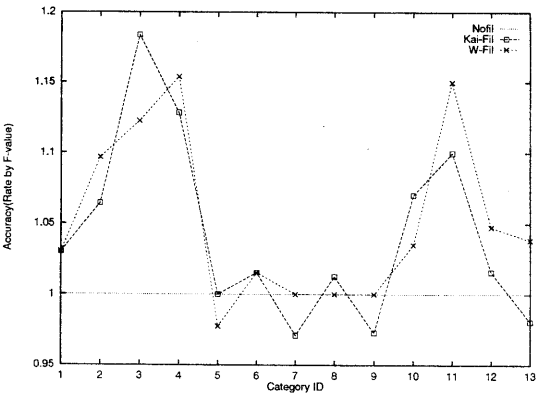

図2: カテゴリー毎のFiltering効果（単語のみ）

図2～4に、実験設定(1)(2)(3)全てに対して、F値に関して分類カテゴリー毎にFilteringをしない場合に

	(1) 単語のみ			(2) 単語＋ALT シソーラス			(3) 単語＋分類語彙表		
	的中率	網羅率	F値	的中率	網羅率	F値	的中率	網羅率	F値
Filtering 無し	70	52	59	74	56	62	68	54	60
χ^2-Filtering	70	55	61	72	60	65	71	55	61
Weight-Filtering	70	55	61	72	61	65	71	57	63

表 3: 実験結果：シソーラス及びFiteringの効果

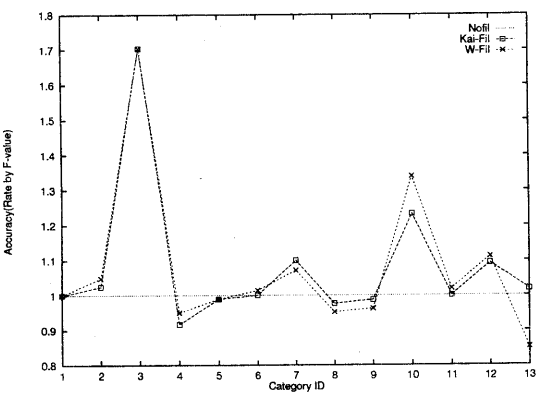

図 3: カテゴリー毎のFiltering効果（単語＋ALT）

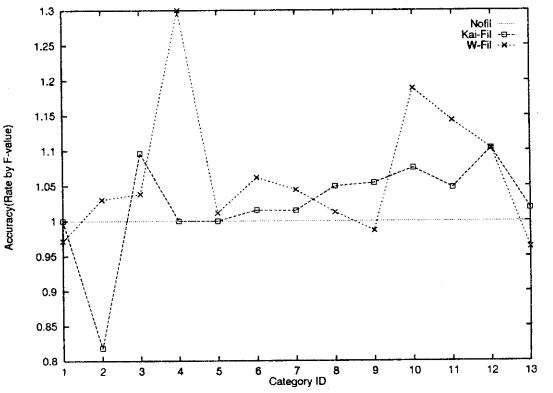

図 4: カテゴリー毎の Filtering 効果（単語＋分類語彙表）

対する、2つの Filtering 手法を用いた場合の精度の相対比 を示す。これらの図から、各実験設定において、幾つかのカテゴリーで Filtering の効果が確認できる。例えば、単語のみを利用した場合（図 2）、特に 3 つのカテゴリー（3. 軍事技術 4. 精密機械 10. サッカー）において、Filtering の効果が大きいことが解る。また、2 種類の Filtering 手法は、どちらも同じ程度の精度向上の効果が得られるので、どちらが良いとは判断できない。

6.3.2 多義性解消の効果

先の実験で得られた、シソーラス利用の場合に精度が劣る 3 つのカテゴリーを多義性解消の対象カテゴリーとして、多義性解消実験を行なった。結果を、図 5、6 に示す。図は、次の 3 つの場合 ((1) 単語のみ (2) 単語＋ALT シソーラス (3) (2)+ 多義性解消）について、3 つのカテゴリーに対する F 値の相対比 ((1) 単語のみの場合を 1 とする）を示す。図で示す通り、χ^2 に基づく手法の場合、3 つのカテゴリー全てで精度が向上し、(1) 単語のみ の場合よりも、精度が良くなることが解る。また、重みに基づく手法は、χ^2 に基づく手法に比べて、効果が小さいことも解った。

以上の実験結果をまとめると

- シソーラスを利用すると、分類精度向上の効果が見られ、ALT シソーラスを利用した場合の方が、より大きな効果が得られる。一方、シソーラス利用で精度が悪くなるカテゴリーも存在する。
- Filtering は、χ^2 値、重みのいずれを利用しても、精度向上効果が得られる。一方、シソーラス利用で精度が悪くなる問題を解決するまでの効果はない。
- χ^2 に基づく多義性解消は、シソーラス利用によって生じる問題を解決する効果がある。

7 まとめ

本稿において、誤り駆動型学習 WINNOW アルゴリズムとシソーラスを利用したテキスト自動分類手法について述べた。

シソーラス利用に伴う問題点として生じる次元数の拡大に対して Filtering 手法を、また多義性の問題に対して分類カテゴリー情報を利用した多義性解消手法を提案し、実験の結果その効果を確認した。Filtering と多

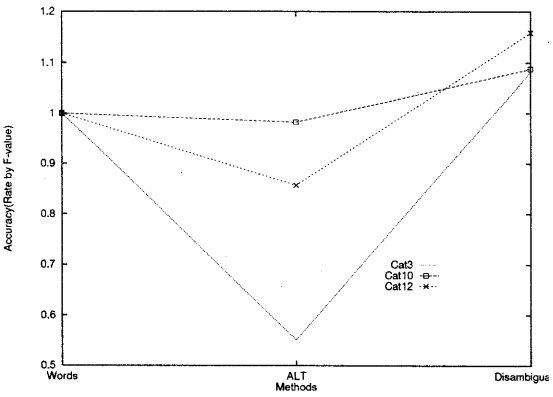

図 5: χ^2 に基づく手法による多義性解消効果

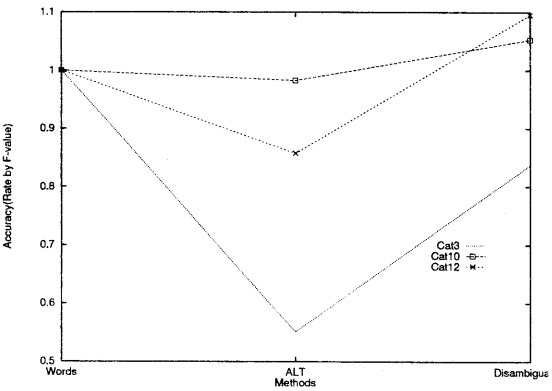

図 6: 重みに基づく手法による多義性解消効果

義性解消により、シソーラスの有効利用が可能になったと考えられる。

今後の拡張の一つの方向として、分類モデルに与える特徴に単語の 2-gram や 3-gram[CS96]、また漢字情報[TI97] を追加することが考えられる。テキストから特徴抽出する際に行なう形態素解析時に、複合語をより細かい単位に誤って分割している場合があり、この誤りが分類精度に影響を及ぼしている問題に対しては、特に単語の 2-gram や 3-gram の利用は有効であろう。また、他の学習手法(例えば、K-近傍法等) との比較も行ない、本手法の有効性をさらに検証する予定である。

謝辞

本研究では、ＣＤ-毎日新聞94年版を利用しました。新聞記事データの研究利用許諾を頂いた毎日新聞社に感謝します。また本稿を作成するに当たって、有益なコメントを頂いた、NTT コミュニケーション科学研究所 春野雅彦研究員に感謝します。

参考文献

[ADW94] C. Apte, F. Damerau, and S. Weiss. Toward language independent automated learning of text categorization models. In *Proc. of the 17th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[Blu95] Avrim Blum. Empirical support for winnow and weighted-majority based algorithms: results on a calendar scheduling domain. In *Proc. 12th International Conference on Machine Learning*, 1995.

[CS96] W. W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. In *Proc. of the 19th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[GR96] Andrew R. Golding and Dan Roth. Applying winnow to context-sensitive spelling correction. In *Proc. 13th International Conference on Machine Learning*, 1996.

[Lew92] D. Lewis. An evaluation of phrasal abd clustered representation on a text categorization problem. In *Proc. of the 15th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.

[LG94] D. Lewis and William A. Gale. A sequencial algorithms for training text classifiers. In *Proc. of the 17th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[Lit88] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, No. 2, pp. 285–318, 1988.

[Lit95] N. Littlestone. Comparing several linear-threshold learning algorithms on tasks involving superfluous attrribute. In *Proc. 12th International Conference on Machine Learning*, pp. 353–361, 1995.

[TI97] Takenobu Tokunaga and Makoto Iwayama. Word-based vs. character-based indexing: An experimental study on japanese text reprsentation for text categotization. In *Workshop in New Challenges in IR and Dissemination*, 1997.

[vR79] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.

[河合 92] 河合敦夫. 意味属性の学習結果にもとづく文書自動分類方式. 情報処理学会論文誌, Vol. 33, No. 9, pp. 1114–1122, 1992.

[国立 64] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[情報 94] 情報科学技術協会. 国際十進分類法. 丸善, 1994.

[池原 93] 池原悟, 宮崎正弘, 横尾昭男. 日英機械翻訳のための意味解析用の知識とその分解能. 情報処理学会論文誌, Vol. 34, No. 8, pp. 1692–1704, 1993.

[豊浦 96] 豊浦潤, 徳永健伸, 井佐原均, 岡隆一. RWCにおける分類コード付きテキストデータベースの開 発. 自然言語処理研究会 NLC 96-13. IEICE, 1996.

[李 97] 李航, 山西健司. 線形結合モデルを用いたドキュメント分類. 自然言語処理研究会 NL 97-119, pp. 37–44, 1997.