自然言語処理 112－14
（1996．3．15）

日本語テキスト分類における特徴素抽出

西野　文人
nisino@flab.fujitsu.co.jp
富士通研究所

テキスト分類や検索においてその文書の内容を表現するインデキシング言語をどのように設定するかは，分類・検索の精度を決定する大きな要因になっている．そこで，特許文書を対象として，特徴素として単漢字，単語，フレーズを使用したインデキシング言語で分類実験を行なうことにより，日本語テキストではどのような特徴素抽出が有効なのかを実験・検討した．その結果，単漢字ベースのインデキシング言語は最も分類精度が悪く，フレーズベースのインデキシング言語が単語ベースのインデキシング言語より良く，単語の bigram モデルによるインデキシング言語が最も良い結果を得た．これらの結果から，多少の雑音の混入を気にせずに有効そうな特徴素を多く抽出することが分類精度向上に良い結果をもたらす可能性が高いということがわかった．

Feature Extraction in Japanese Text Categorization

Fumihito NISHINO

Fujitsu Laboratories Ltd.

The objective of this paper is to overview some indexing languages. We present some results of the patent categorization experiments. These experiments indicate that the word-based indexing language is better than the single-kanji-based indexing language, and the phrasal indexing language is better than word-based indexing language. In these experiments, the indexing language by word-based model obtained the best result.

1 はじめに

テキスト分類や検索では，テキストからどのような特徴素を抽出して，そのテキストの内容を表現したインデキシング言語を作成するかが問題となる．英語を対象とした分類・検索では，特徴素として単語を使用した単語ベースのインデキシング言語が使われることが多い．これは英語では単語という単位が明確であり簡単に切り出せるということによるところが大きい．これに対して日本語を対象とした分類・検索では，日本語の単語という単位が不明確であることや単語を切り出すコストが大きいというような問題点がある．そこで，特徴素の抽出が簡単であり，特徴素数の上限も高々数千に限定されている，というような取り扱いやすさの点から単漢字を特徴素として利用した分類実験も行なわれている [1]．しかし，カタカナ語を取り扱わないことを含め，形態素解析のような自然言語処理技術を利用しない単純な手法のための精度低下がどの程度のものなのかそこには報告がなく大きな疑問として残る．

日本語に限らず，自然言語処理解析技術を活用して良い特徴素を抽出したら検索・分類の精度があがるのではないかという期待もある．このような試みの一つとして，Lewis はフレーズをインデキシング言語として使用してロイター通信の全文ニュース記事を対象に分類実験を行ない，単語ベースのインデキシング言語とその分類精度を比較するという実験を行なっている [2]．しかしながら，この実験の結果はフレーズインデキシング言語を使った分類は単語ベースのインデキシング言語に比べてかなり低い結果となっている．だが，この実験では，英語を対象とした実験であるということ，および，ここでの実験におけるフレーズインデキシング言語としては単一の名詞からなるフレーズは取り除かれているということ，訓練文書数をもっと多くしてやったらどうなったのかが不明なことなど，それらの環境が少し違えば分類の結果もまた異なったものになる可能性もあり，この報告だけからフレーズインデキシング言語が性能が良くないと結論づけるわけにはいかない．

テキスト分類では，単に訓練文書数や分類項目数の数だけでなく，分類の粗さや分類項目のまとまり具合によってその分類精度は大きく変化する．したがって，手法の善し悪しを議論するには絶対的な分類精度はほとんど意味をなさない．そこでインデキシング言語の違い（特徴素抽出の違い）によるテキスト分類精度の違いを特許文書を分類して国際特許分類コードを付与する実験を通して調べた．

以下では，まずこの実験に先だって開発した日本語文書から名詞句を抽出する処理について述べ，その後，単漢字，単語，フレーズのそれぞれを特徴素としたインデキシング言語に対する特許分類実験結果について述べ，考察を行なう．

2 名詞句抽出

日本語テキストの分類や検索においてはその文書の内容を表現したインデキシング言語を構成する特徴素を抽出することが必要となる．その特徴素抽出として単漢字や n-gram といった自然言語処理解析技術を特に必要としない方法もあるが，最近では形態素解析を利用した特徴素抽出も手軽に行なえるようになってきた．しかし，汎用の形態素解析プログラムの解析結果をそのまま利用した単語単位のタームをインデキシング言語としようとした場合，専門用語として一つにまとまるべきものが形態素単位に細かく分割されすぎたり，形態素解析の能力不足などから専門用語がうまく抽出されないことも多い．さらに，分野別に分類するための特徴素や検索のためのキータームとしては不要な単語も取り出してしまうなどの問題もある．そこで，テキスト分類や検索のインデキシング言語となる特徴素を抽出するために，単に形態素解析をしてテキストを形態素に分割するだけでなく，各形態素の意味的な役割も少し考え，特徴素となり得る可能性の高いフレーズ（名詞句）を抽出するルーチンを作成した．図 1は特許文書に対して本名詞句抽出プログラムを利用して名詞句を抽出した例（抽出したものは【】で括弧づけされている）である．このプログラムの特徴としては以下のことをあげることができる．

1. 名詞連続の抽出

   「比例流量制御バルブ」，「対地作業機」のような名詞連続（接頭語，接尾語を含む）は一つの名詞句として抽出する．このような専門用語は通常辞書に未登録であり，形態素解析の結果としては変な形態素の区切り方になっていたとしても名詞句としての範囲の認定は比較的うまくいく．

2. 動詞連用形の処理

   和語動詞は抽出の対象としていないが，「ばらつき」のような連用形で表現され前後の状況などから名詞的に使用されていると判断される場合には抽出する．

3. 名まえの抽出

   「ベイズの定理」「マーフィーの法則」などの名まえは連体助詞「の」を含めて抽出する．

4. 状態性名詞の除外

   「機械的」や「適正」，「一定」のような状態を示すものが判定辞に前接している場合にはこれを抽出しない．しかし名詞連続中に出現する場合には抽出する．例えば，「xx が異常だ」の「異常」は xx の状態を示すものとして抽出しないが，「異常気象」のような使われ方の場合には「異常気象」というタームとして抽出する．

【比例流量制御バルブ】の機械的な【ばらつき】や【油温】の【変化】を考慮し、適正な【駆動オンタイム】の【補正】が行われるようにする。一定の【速度測定区間】を設け、予め【規定】した【オンタイム】で【規定時間】だけ【比例流量制御バルブ】を【駆動】して【速度】を【測定】する。そして、【規定】した【速度】との【差】に応じて【駆動オンタイム】を【補正】するが、【制御】中の【セットオンタイム】が【規定値】以上のときのみ前記【駆動オンタイム】の【補正】を行うように【制御】する。【比例流量制御バルブ】により【対地作業機】の【昇降制御】を行う【動力農機】であって、予め【規定】した【オンタイム】で【比例流量制御バルブ】を【規定時間駆動】したときの【変位】を【測定】し、予め【規定】した【速度】と実際に【測定】された【速度】との【差】に応じて【駆動オンタイム】を【補正】する【油圧制御装置】に於いて、【制御】中に【セット】された【セットオンタイム】が【規定値】以上のときのみ、前記【駆動オンタイム】の【補正】が行われるようにしたことを特徴とする【動力農機】の【油圧制御装置】。

図 1: 特許文からのキーフレーズ抽出

5. 分野性のない名詞の除外

   文書の分野性を示さない名詞（文書中の構造や他の場所を指定したり，筆者の思考や心的状態を示したり，事象間の関係などを示すようなもの）は抽出しない．

6. 連体詞的，相対的，副詞的なものの除外

   「該」，「同」や，「中」，「付近」，「以上」，「現在」など，連体詞的，相対的，副詞的なものは名詞句の一部としない．

7. 番号の除外

   「レバー 3」，「センサ 4」などの名詞連続の後の数字（特許文書では頻出する）はこれを取り除く．これらは前の名詞の単なる ID と考えられるからである．しかし，「号」や「世」など特定の接辞とともに用いられる数字はこれを含めて抽出する．

8. アルファベット

   「DOPI」，「FM」のような 2 文字以上からなるアルファベット列はこれを抽出する．1 文字のアルファベットは記号と考えて抽出しない．

3 自動分類実験

特徴素の違いの影響を，ここでは特許の文書を対象に国際特許分類コードを付与する実験で調べた．類似度を計算する手法としては，文書 d がカテゴリ C に含まれる確率を特徴素 t_i に対して，

$$P(d \in C|D = d) = P(d \in C) \sum_{t_i} \frac{P(T = t_i|d \in C)P(T = t_i|D = d)}{P(T = t_i)}$$

とした岩山の提案した SVMV 法 [3] を用いた．また，カテゴリ割り付けには文献 [4] で示した信頼度に基づく方法（確率モデルに基づいて計算した $P(d \in C|D = d)$ の値が p 以上であった時に実際に文書 d がカテゴリ C に属する確率（つまり正解である確率：信頼度）を訓練文書から推定し，p と信頼度との関係に基づいてカテゴリ割り付けを決定する方法）を用いた．分類の粗さとしては特許分類の一番上位のセクションによる 8 分類（A: 生活必需品，B: 処理操作；運輸，C: 化学；冶金，D: 繊維；紙，E: 固定構造物，F: 機械工学；照明；加熱；武器；爆破，G: 物理学，H: 電気）で実験した．訓練文書としては特開平 5-038201 から特開平 5-077241 までの 39,000 件を，テスト文書としては．特開平 5-077242 から特開平 5-084000 までの 6749 件を用いた[1]．

文書から何を特徴素として抽出するかというインデキシング言語に関しては，以下のもので比較した．

1. 単漢字 (KC)

   文書中の全漢字を一文字単位に取り出したものをインデキシング言語とした．

2. 名詞単漢字 (NKC)

   全漢字を取り出したのでは「該」とか「以上」のような分野の特徴とは無関係な単語における漢字からもそれらの漢字が特徴素として取り出されることになる．これらは各特許に平均的に出現するだろうから重みが小さくなってあまり影響はないものと考えることもできるが，これらの雑音の影響を調べるために，まず形態素解析を行なって名詞単語だけを抽出し，抽出された名詞単語に対して全漢字を一文字単位に取り出してインデキシング言語とした．

3. 漢字単語 (KW)

   単漢字インデキシング言語の問題はカタカナ語やひらがな語を特徴素として抽出しないことにもある．そこで，形態素解析を行なって名詞単語だけを抽出し，その中から漢字のみからなる単語だけを特徴素として残したものをインデキシング言語とた．すなわち，通常の単語インデキシング言語

[1] これらの条件は文献 [4] と同じにしてある

と比較して，ひらがなやカタカナを含む名詞単語をインデキシング言語から除外したということである．

4. 単語 (W)

   形態素解析を行なって得られた名詞単語をインデキシング言語とした．

5. フレーズ (P)

   名詞句抽出プログラムによって抽出したフレーズ（名詞句）をインデキシング言語としたもの．

6. 複合語 (P-W)

   Lewis[2] でのフレーズインデキシング言語では単一の語からなる名詞句はインデキシング言語から取り除いている．そこで，名詞句抽出プログラムによってフレーズを抽出するが，抽出したフレーズが単一の単語からなっているときはそれを取り除き，複合語だけをインデキシング言語とした．

7. 漢字 bigram(KC2)

   テキスト中の漢字の 2 文字連続，いわゆる漢字 bigram をインデキシング言語とした．単独で出現する漢字や，カタカナ，ひらがな，アルファベットは無視した．例えば「日本語テキスト分類」からは「日本」，「本語」，「分類」を特徴素として抽出することになる．

8. 名詞句漢字 bigram(PKC2)

   単純な漢字 bigram では，「以上」や「上述」といった副詞的名詞，記述性名詞などの分野性のない単語からも漢字 bigram データが取られる．そこで，これらの雑音を取り除いた効果を調べるために，最初に名詞句抽出プログラムで名詞句を抽出し，その中だけで漢字 bigram を取ったものをインデキシング言語とした．

9. 単語 bigram(W2)

   名詞句を抽出して，それが 2 形態素以上からなる時，あらゆる 2 連続形態素を特徴素として抽出してインデキシング言語としたもの．例えば，「日本語テキスト分類」は，「日本語」，「テキスト」，「分類」の 3 形態素に分割されるが，ここからは，「日本語テキスト」と「テキスト分類」の 2 つを特徴素として抽出することになる．

4 実験結果

4.1 各インデキシング言語における特徴素数と特徴素抽出速度

まず，39,000 の訓練文書から特徴素を抽出した時の各インデキシング言語での異なり特徴素数およびテスト文書における 1 件あたりの特徴素数を表 1に示す．テスト文書 1 件当たりの特徴素数は 100 ～ 300 の間におさまっているのに対して，訓練文書中での異なり特徴素数は単漢字インデキシング言語では 3000 強に対してフレーズインデキシング言語では 150 万弱と大きな差を生じている．この差は，各特徴素に対して同一性をチェックして頻度をカウントしたり，出現する文書 ID を保持しておくなどの管理の手間において大きな差を生じることになる．

インデキシング言語	訓練文書中の異なり特徴素数	テスト文書 1 件当たりの特徴素数
単漢字	3429	245
名詞単漢字	3353	188
漢字単語	35093	153
単語	259309	224
フレーズ	1465542	227
複合語	1309138	108
漢字 bigram	217987	261
名詞句漢字 bigram	187802	197
単語 bigram	910157	127

表 1: 各インデキシング言語の特徴素数

特徴素抽出の速度は，漢字 bigram を取るのに時速 10Gbytes に対して，前述のプログラムを利用して形態素解析を行なって名詞を抽出するあるいは名詞句を抽出すると時速 120Mbytes と約 80 倍かかる．しかし，これは特徴素を抽出するだけの時間で，実際には抽出された特徴素の同一性のチェックをして頻度をカウントするなどの時間がかかることになるので，例えば，この部分の速度が時速約 600Mbytes とすると（同一性のチェックなどに単純なソートコマンドを利用するとこの数倍以上はかかる），この時間まで含めて考えれば，単漢字や n-gram などの形態素解析を行なわない特徴素抽出手法と，形態素解析を行なって特徴素を抽出する手法との速度比は 6:1 程度ということになる．ちなみに特許文書テキストは 1 万件あたり約 140Mbytes である．

4.2 インデキシング言語と分類精度

前セクションで示したそれぞれのインデキシング言語に対する特許分類実験結果を図 2に示す．

単漢字インデキシング言語

まず，この中では単漢字を使用したインデキシング言語の分類精度が最も悪い．例えば適合率 (precision) を 80% 水準で比較してみると，単漢字の精度は，形態素解析を利用した単語ベースのものと比較して 30 ポイントも悪いことになる．

単漢字インデキシング言語に比べると，形態素解析によって名詞を抽出してからの単漢字インデキシングは

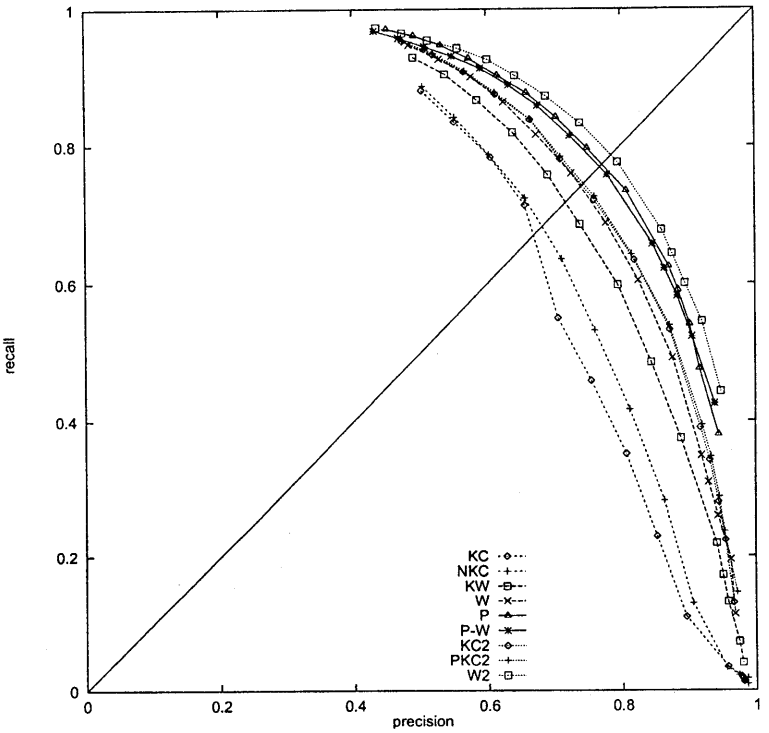

図 2: 各種インデキシング言語による粗分類

わずかな精度向上がみられる．これは単漢字では抽出されてしまった非名詞を形成する漢字文字が取り除かれた結果といえるが，テスト文書 1 件当たりの特徴素数の差と訓練文書中の異なり特徴素数の差から考えると，単に雑音となる文字が取り除かれたというより，ある文字が分野分類に意味のある使われ方の時と分野分類には意味のない使われ方の時とがあって，非名詞に対する雑音を取り除くことで意味のある使われ方の時の特徴がはっきりと目立つようになったものと考えるのが妥当であろう．

名詞単漢字と漢字単語ではほぼ同じ部分を特徴素範囲として抽出しているが，単一文字を特徴素とするか単語そのものを特徴素とするかの違いがある．しかし，分類結果を見ると単語を特徴素とする方が良い分類情報が得られている．すなわち，文字ベースより単語ベースの方が分類精度が良いということである．

単語インデキシング言語に比べて，漢字単語インデキシング言語は，分類精度が落ちている．このことは，非漢字部分（特にカタカナ語と考えられる）を特徴素として加えることが有効であることを示している．

以上から，単漢字ベースのインデキシング言語は，特徴素の抽出は非常に簡単であり，また特徴素数の増加も少ないので取り扱いは非常に楽であるが，単語と文字の特徴素の粗さの違い，非名詞部分の雑音の影響，非漢字（特にカタカナ語）を取り扱っていない点などから，分類精度の面では問題が多い．

フレーズインデキシング言語

単語インデキシング言語に比べて，フレーズインデキシング言語は，1 文書あたりの特徴素数はほとんど変わらないが，39000 件の訓練文書全体で見た異なり特徴素数は単語インデキシング言語で 26 万特徴素なのに対してフレーズインデキシング言語では 147 万特徴素と 5.7 倍もの開きがある．分類精度は，Lewis の結果とは異なり，フレーズインデキシング言語の方が優位であり，フレーズの特徴素が有効に働いているようである．これは，詳細な分類ではいざ知らず，特許の粗分類程度の粗さの分類でもこのような結果がでたというのは意外であった．

特徴素として，単一の語からなるフレーズを除いたものと除かなかったものとでは特に差がなかった．前示した図 1の結果をみると，名詞句抽出ではもっと知的な処理を加えて，さらに不要そうな単語は抽出しないようにすることが必要そうにも思えるが，この結果を見るかぎり，そういった処理は特に不要かもしれない．

漢字 bigram インデキシング言語

漢字 bigram はその抽出のしやすさという点では単漢字に匹敵するが（しかし，特徴素数では単語と同じくらい増加する），分類精度は単漢字に比べるとはるかに良く単語インデキシング言語とほぼ同等である．非漢字部分を扱っていない（漢字単語インデキシング言語と比較して優位である）ことを考えると，それらを扱う処理をうまく導入すること（特にカタカナが問題になるが，カタカナは bigram をとってもあまり意味がなさそうなので，例えば漢字 bigram に付け加えてカタカナ文字列をインデキシング言語とするなど）でさらに良い結果を得られることが期待できる．

単語 bigram インデキシング言語

フレーズインデキシング言語では「日本語テキスト分類」と「英語テキスト分類」とでは全く違った特徴素として取り扱われるが，単語 bigram では「テキスト分類」が共通の特徴素として扱われる．またフレーズインデキシング言語では単一語からなる名詞句を特徴素から除いても精度に変化がないという結果も合わせて単語 bigram インデキシング言語は適度な特徴素数と頻度が期待された．

結果は図 2で明らかなように，今回行なった一連の実験の中では最も良い分類精度を得ることができた．特徴素数を見てもフレーズインデキシング言語より少なく，いくらか扱いやすいものになっている．

4.3 訓練文書数の影響

訓練文書数 39000 が十分な訓練量であったのかどうかを確認するために，単漢字，単語，フレーズのそれぞれを特徴素とするインデキシング言語に対して，訓練文書を 5000，10000，20000，30000 とした場合の分類実験を行なった．それぞれの訓練文書数での特徴素数を表 2に，分類実験の結果を図 3に示す．

単漢字インデキシング言語では訓練文書数が 20000 までは分類精度の向上が見られるが，その後は訓練文書がいくら増えても分類精度にほとんど変化はみられない．39000 件の訓練文書は十分な量の文書数と考えることができる．しかし，その精度は訓練文書が 5000 のフレーズインデキシング言語にも単語インデキシング言語にも劣るものであった．

単語インデキシング言語もやはり訓練文書数が 20000 までは分類精度の向上が見られるが，それ以上は訓練文書の数が増加しても分類精度にほとんど変化はみられない．訓練文書数 20000 での異なり特徴素数から訓練文書数 39000 の異なり単語数では，約 1.6 倍の増加があるものの，これらはほとんどが分類精度に影響を与えない非重要語といえよう．

フレーズインデキシング言語では，訓練文書数が 5000 件の場合には，単語インデキシング言語と同程度の分類

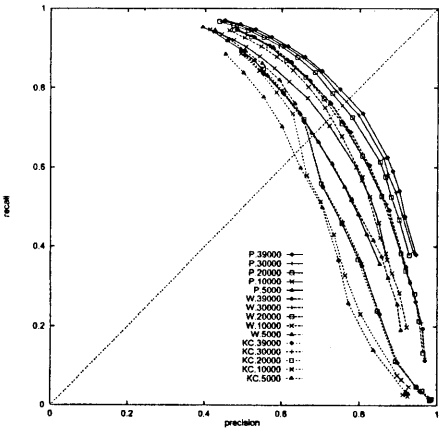

図 3: 訓練文書数と分類精度

精度しか出ていないが，訓練文書数が 10000 以上では，単語インデキシング言語より良い分類精度を示している．そして，単漢字インデキシング言語や単語インデキシング言語とは異なり，訓練文書数が 30000 から 39000 に増加してもまだ分類精度が向上している．これは，フレーズインデキシング言語では，訓練文書 39000 は十分な数とは言えず，訓練文書を増やすことによって，まだ精度がよくなる可能性があることを示している．

4.4 低頻度語の除去

文書頻度（いくつの文書に出現したか）が低い特徴素は雑音である可能性が高い．そこでフレーズインデキシング言語において，文書頻度 (df) が 1 以上（すなわち全ターム）を特徴素としたものと，文書頻度 2 以上（すなわち文書頻度 1 のタームを特徴素として取らなかった）だけを特徴素とした場合，同様に，文書頻度 3 以上，5 以上，9 以上，17 以上のもので分類実験を行なった．

表 3は文書頻度 n 以上の特徴素数と全体に占める割合を示したものである．この表から，フレーズインデキシング言語では，文書頻度 1 の特徴素が 70% 以上を占めることがわかる．

文書頻度	特徴素数	割合
1 以上	1465542	1
2 以上	422497	0.29
3 以上	251045	0.17
5 以上	141053	0.096
9 以上	77980	0.053
17 以上	42839	0.029

表 3: 文書頻度ターム（フレーズ）

それぞれに対して分類実験を行なった結果を図 4に

インデキシング言語 ＼訓練文書数	5000	10000	20000	30000	39000
単漢字	2468	2833	3071	3303	3429
単語	70716	108341	158921	212979	259309
フレーズ	301888	509350	849470	1179653	1465542

表 2: 訓練文書数と異なり特徴素数

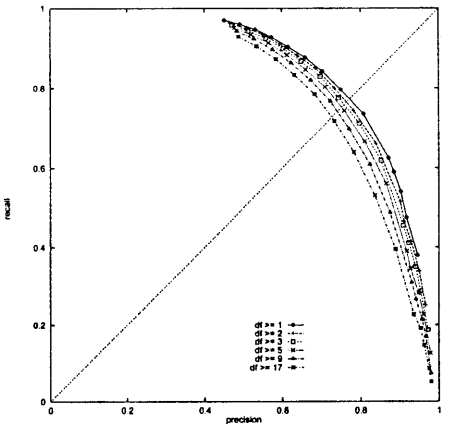

図 4: 低頻度語の除去（フレーズ）

示す．文書頻度の低い特徴素を取り除くと結果が悪くなっている．これはフレーズインデキシング言語の場合，訓練文書数が 39000 ではまだ不十分で全体の 70% 以上を占める文書頻度 1 の特徴素の中にまだ重要な特徴素が含まれていることを意味していると考えられる．

単語インデキシング言語でもフレーズインデキシング言語の実験と同様に文書頻度の低い特徴素を取り除いて実験を行なってみた（表 4，図 5）．こちらは，文書頻度 1 の特徴素を取り除いても分類精度にほとんど変化はなかった．すなわち文書頻度 1 の特徴素には重要な特徴素は少ないということ，そして雑音として特に悪さをするわけではないことがわかる．さらに文書頻度 16 以下のものを取り除いて全体の 8% の特徴素だけを使って分類を行なってもその精度はほとんど変わらず，有効な特徴素は見た目よりかなり少ないことがわかる．

これから訓練文書数 39000 というのは特許のセクション分類程度の粗さならば単語インデキシング言語では，重要な特徴素は複数の文書に出現するくらいの訓練文書であったということを意味すると考えられる．

漢字 bigram インデキシング言語に対する文書頻度と特徴素の数と割合を表 5に，またそれぞれに対する分類実験結果を図 6に示す．漢字 bigram によるインデキシング言語も単語インデキシング言語と同様に，訓練文書 39000 で特徴素は既に飽和しており，この特徴素抽出方法ではこれ以上の分類精度の向上は望めなさそうである．

文書頻度	特徴素数	割合
1 以上	259309	1
2 以上	104628	0.40
3 以上	72036	0.28
5 以上	47315	0.18
9 以上	31137	0.12
17 以上	20798	0.08

表 4: 文書頻度ターム（単語）

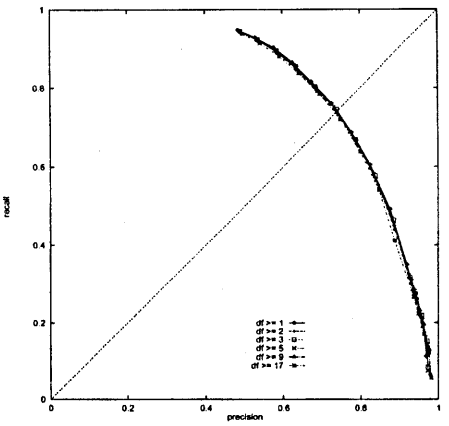

図 5: 低頻度語の除去（単語）

文書頻度	特徴素数	割合
1 以上	217987	1
2 以上	136823	0.63
3 以上	106628	0.49
5 以上	78411	0.36
9 以上	55016	0.25
17 以上	36978	0.17

表 5: 文書頻度ターム（漢字 bigram）

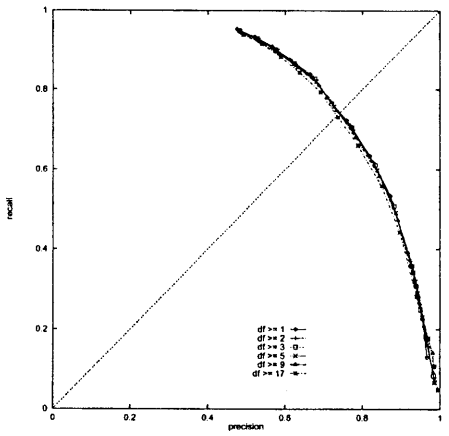

図 6: 低頻度語の除去（漢字 bigram）

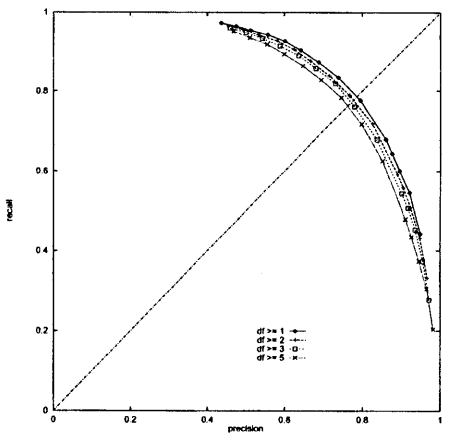

図 7: 低頻度語除去（単語 bigram）

単語 bigram インデキシング言語に対する文書頻度と特徴素の数と割合を表 6に，またそれぞれに対する分類実験結果を図 7に示す．文書頻度 1 の特徴素の割合は単語インデキシング言語とほぼ同じ 61% であるが，こちらは文書頻度 1 の特徴素を取り除くと，フレーズインデキシング言語と同様に分類精度の悪化が見られる．これは，フレーズインデキシング言語と同様にまだ訓練文書数が十分ではなく，訓練文書数を増加させることによってさらなる分類精度の向上の可能性があることを示しているといえよう．

文書頻度	特徴素数	割合
1 以上	910157	1
2 以上	354358	0.39
3 以上	227112	0.25
5 以上	134479	0.15
9 以上	75775	0.08

表 6: 文書頻度ターム（単語 bigram）

5 おわりに

ここではインデキシング言語の違いによるテキスト分類の精度の違いを特許文書の粗分類（セクション分類）の実験結果として示した．この実験から単漢字のベースのインデキシング言語では特徴素の数が少なすぎて特許の粗分類程度の粗さの分類でも十分な精度が得られないことが確認された．一方，漢字 bigram によるインデキシング言語は特徴素抽出が高速に可能であり，しかも非漢字部分の処理をうまく取り込むことでさらに精度向上が得られる可能性があり，有望なインデキシング言語の一つである．そして，純粋に分類精度だけからすれば，訓練文書が十分にあるならばフレーズインデキシング言語と単語 bigram インデキシング言語が有望である（今回の実験では単語 bigram が優勢であったが，どちらもまだ訓練文書を増やせば分類精度向上が望まれ，この優位性はどこかで逆転するかもしれない）．しかし特徴素の数は非常に多くなり，それをどう取り扱うかが問題となる．

以上の実験結果をまとめると，雑音は統計的に無視されるので，多少の雑音が混入することは気にせずに，有効そうな特徴素数の量を増やすことが分類精度の向上に貢献すると言えそうである．

今回の実験は特許文書の粗分類ということでおこなったが，特許分類でも詳細分類や，固有名詞の多い新聞記事の分類では違った結果が得られる可能性もあるが，今回の実験はこのようなシステムを作成する上でも一つの目安となるであろう．

参考文献

[1] 渡辺靖彦，竹内雅人，村田真樹，長尾眞：χ^2 法を用いた重要漢字の自動抽出と文献の自動分類，信学技報，NLC94-25, pp. 23–30 (1994).

[2] Lewis, D. D.: An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task, in *Proc. of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–50 (1992).

[3] 岩山真，徳永健伸：自動文書分類のための新しい確率モデル，情処研報，FI33-9, pp. 47–52 (1994).

[4] 西野文人：テキスト分類のためのカテゴリ割り付け戦略，情処研報，NL106-3, pp. 13–18 (1995).