相互情報量を用いた単語の分類の詳細化

柏岡 秀紀　　Ezra W. Black
ATR音声翻訳通信研究所

1 はじめに

単語の分類体系は，形態素解析や構文解析等で用いられる重要な知識の一つであり，品詞による分類や意味概念による分類等が考えられる．分類体系を構築する一つの手法として，テキストにおける単語間の相互情報量を分類基準に用いた手法[1]が提案されている．この手法により得られる分類体系は，二分木により表現される．この手法では，処理コストを下げるための工夫がなされている．しかし，その工夫により分類対象となる単語が増加すると，単語の処理順序に依存した分類になり，得られる二分木に単語を識別するためだけの冗長な部分木を含んでしまう．

本稿では，Brown等の手法に再分類の機構を加えることにより，この冗長な部分をより詳細な分類体系を表現する部分木として分類体系を得る手法を提案する．再分類する手法として，まとめられた単語を二つずつまとめていく手法と，二つに分割していく手法について議論する．また，分類基準として相互情報量以外の基準を考慮し，各々の手法の特徴を調べ，その実験結果を示す．

2 相互情報量を用いた分類

単語の分類基準として2-gramに基づく相互情報量を用いた処理は，以下のような手順で分類される．

1. すべての単語に対して，一つのクラスを割り当てる．
2. 全てのクラスの中から，可能な二つのクラスの組合せのうち相互情報量の減少を最小にする組合せを選択し，この二つのクラスをまとめて新たな一つのクラスを作成する．
3. 全ての単語が一つのクラスになるまで2に戻る．

上記のアルゴリズムでは，単語数(語彙数)に依存して記憶領域と計算量が非常に大きくなる．現実的な処理を行なうためにBrown等は，組み合わせる処理対象をある一定数のクラスに限定し，分類基準としての相互情報量の計算処理にかかる計算量を抑える手法を提案している．これは，上記のアルゴリズムにおける2の処理を，全てのクラスの中からまとめるクラスを選択するのではなく，一定数N個のクラスから選択する事で実現している．

この手法を実現するために上記の処理の流れを三つに分割し，各々の処理を以下のように行なう．

1. 単語の処理順序の決定
   - 単語の出現頻度を調べる．
   - 出現頻度の高い単語から順に，一定数N個を分類対象として次の処理に移る．
2. 一定数のクラスでの分類
   - 全ての単語がN個のクラスに割り当てられるまで以下の処理を繰り返す．
     (a) N個のクラスにおいて，可能な二つのクラスの組合せのうち相互情報量の減少を最小にする組合せを選択する．
     (b) 選択された二つのクラスをまとめて新たな一つのクラスを作成する．[1]
     (c) 分類対象外の単語の中から，出現頻度の高いものを新たにN番めのクラスとして分類対象に加える．
3. 一定数のクラスをまとめる
   - 1個のクラスにまとまるまで以下の処理を繰り返す．
     (a) 現在あるクラスにおいて，可能な二つのクラスの組合せのうち相互情報量の減少を最小にする組合せを選択する．
     (b) 選択された二つのクラスをまとめて新たな一つのクラスを作成する．

[1] これによりN個のクラスが$(N-1)$個のクラスになる．

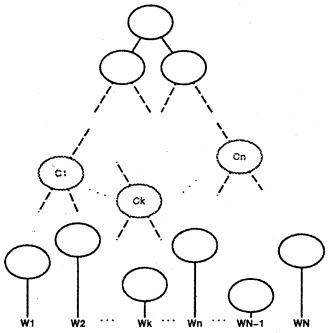

図 1: 理想的に得られる分類体系

上記の処理で二つのクラスは，これら二つのクラスをサブクラスとして持つクラスを作成することによりまとめる．全てのクラスは二つのサブクラスから構成されるために，最終的な分類体系は，二分木として表現され(図1)，単語は二分木の葉にすべて割り振られる．

分類実験

この手法により the AP Newswire のテキストを用いて実験を行なった．一月分のテキスト(総語彙数：約 75,000 語，総語数：約 3,800,000 語)のうち，出現頻度が 10 より多い約 16,000 語彙と，一年分のテキスト(総語彙数：約 300,000 語，総語数：約 40,000,000 語)のうち，出現頻度が 10 より多い約 60,000 語彙について行なった．また，short abstracts from the Department of Energy のテキストに対しても実験を行なった．

各実験において処理2における一定のクラス数は，500 として処理を行なった．どの実験においても，一定数の 500 クラスでの分類の処理において，最初の約 1500 単語(一クラス平均 3 単語)を処理した後は，新たに分類対象となった語のクラスが，それまでに対象となっていたクラスとまとまるだけであった．まとめ方が，「まとめられる二つのクラスをサブクラスとして取る新たなクラスを作成する．」という手法を取っているために，一定数の 500 クラスの内部を表現する木構造が，非常に偏ったものになる(図2)．最終的な木構造は，この 500 クラスをまとめあげることにより得られるため，分類体系としては，不適切な部分木を含む構造になる．

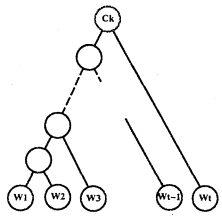

図 2: 偏った部分木

この 500 クラスの部分木を適切な構造にするために，単語の分類順序を変更する手法や二つのクラスをまとめ方を変更する手法，500 クラスの部分木を各々再構成する手法などが考えられる．以下では，部分木を再構成する手法について議論する，

3 再分類による詳細化

相互情報量による一定数のクラスへの分類では，500 クラスの各部分木に，単語の分類順序による偏りが生じている．各クラス毎に属している単語を対象として再分類することで，分類順序による偏りを少なくし，より詳細な分類が得られると思われる(図3)．再分類の分類基準としては，一貫した分類体系を得るために用いる相互情報量と，他の分類基準についても検討する，再分類では，対象となる語の中から分類基準により語を二個ずつまとめあげていく手法と，対象となる語を二つに分割していく手法がある．ここでは，各手法の計算量等の特徴を調べる．

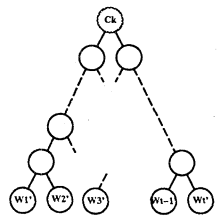

図 3: 再分類後の部分木

3.1 分割していく手法

一定数の500クラスの各クラスの再分類では，対象となる語が平均して全体の語彙数の500分の1になる．あるクラスでの再分類の処理対象をVとすると，このクラスを，一回分割するために調べる必要のある分割の組み合わせは，2の$V-1$乗通りあり非常に大きな計算量を必要とする．再分類の目的である分類体系の偏りを減らすことが，左右の部分木に同程度の数の語があるようにする事だと考えると，分割するために調べる必要のある組み合わせは，かなり減少する．

3.2 まとめあげていく手法

この手法は，一定数のクラスをまとめる処理と同様に行なうことができる．分類する処理の対象の語彙数をVとすると，一つの新しいクラスを作るために，まとめる二つの語の組合せを調べるために，$O(V^2)$の処理が必要となる，Vがある一定数をこえると計算量が大きくなり，一定数のクラスに限定することにより計算量を抑える必要が出てくる．しかしこの場合には，分類順序に依存した問題が再分類においても同様に生じることになる．この場合には，再分類した結果に対して，さらに再分類を行うことで問題を解消することが可能と思われる．

再分類に用いる分類基準として相互情報量を用いることは，得られる分類体系に一貫した基準を与えるという点で良い基準と思われるが，同じクラスに属する語は，一度，相互情報量の基準により分類されている点と，各語の出現頻度に左右されることから，他の基準を用いた実験も行なった．他の分類基準として，次のような基準について調べてみた．以下に示す式において，$C()$は，出現頻度を示している．

同じ単語が隣接している割合による基準

この基準は，以下の式により計算される．

$$\sum_{w0}(min(C(w0,c1),C(w0,c2)))/(C(c1)+C(c2)) + \sum_{w0}(min(C(c1,w0),C(c2,w0)))/(C(c1)+C(c2))$$

ある単語w0が，二つのクラスc1の前(後)とクラスc2の前(後)に隣接して現れる頻度の小さい方と，クラスc1,c2の頻度の和の割合を計算する．単語w0を全ての語に対して計算し，和を取ったものである．

同じ単語が隣接している確率による基準

この基準は，以下の式により計算される．

$$\sum_{w0}((C(w0,c1)/C(w0))(C(w0,c2)/C(w0))) + \sum_{w0}((C(c1,w0)/C(w0))(C(c2,w0)/C(w0)))$$

上記の二つの基準は，相互情報量の基準と良く似た考え方に基づいていて，「関連性の強い単語から順にまとめられる」という特徴を持っている．これらの基準では，最も高い値をとるクラスc1，c2の組合せに対して，この二つのクラスをサブクラスとして持つ新たなクラスを作ることで，まとめあげていく．したがって，相互情報量による分類と同様に分類結果は，二分木で表現される．

4 再分類の実験

再分類の実験には，2節で示した the AP Newswire の分類結果を用いた．一月分のテキスト約16,000語彙の場合，一定数500の各クラスは，一クラス平均約30語[2]であった．本節では，次のように分類されたクラス(図4)を例に実験結果を示す．

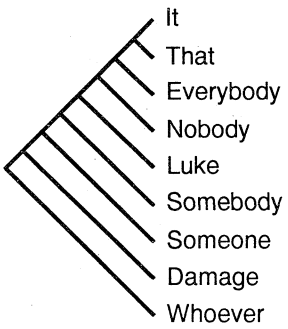

図 4: 再分類前の部分木

[2]実験結果では，最大1145語，最小1語のクラスができている

4.1 分割による再分類

平均30語のクラスを一回分割する組合せは，約10億(2の29乗)通りあり，実際の分類処理には不適切と思われるために，今回は実験を行っていない．

4.2 まとめあげによる再分類

まとめあげていく手法では，一回にまとめることのできる組合せは，一クラス平均30語とすると，約500通りとなる．ここでは，先に述べた同じ単語が隣接している割合による基準，および同じ単語が隣接している確率による基準による再分類結果を示す．

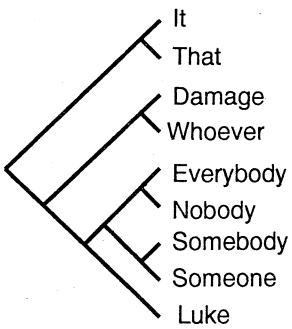

図5: 同じ単語が隣接している割合による部分木

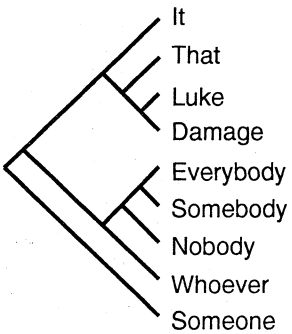

図6: 同じ単語が隣接している確率による部分木

これらの再分類された部分木を見ると，再分類前の部分木に比べ，詳細な分類が得られていると思われる．例えば，このクラスをより細かなクラスとしてとらえようとすると，rootノードより分割されている部分木を取り出すことになる．再分類前の部分木では，“Whoever”，“Damage”の順に一単語ずつ分けられるが，図5では，“It, That”とそれ以外，さらに“Damage, Whoever”とそれ以外のようになる．図6では，“Someone”とそれ以外，さらに，“It, That, Luke, Damage”と“Everybody, Somebody, Nobody, Whoever”のように分けられている．つまり，再分類する処理は，すでに得られている分類体系をより詳細な体系としており，各クラス内に構成されている二分木をより有効に構成するという利点を持つといえる．

5 おわりに

相互情報量を用いた単語の分類について述べ，これまでに提案されている手法の問題点として，得られる分類体系の分類順序による偏りを示し，実際のテキストデータに対する分類実験により明らかにした．この偏りを修正する手段として再分類する手法を提案し，幾つかの新たな分類基準により再分類の実験を行った．その結果，部分木の偏りを解消し，より詳細な分類体系が得られる事を確認した．

今後は，相互情報量を用いた分類手法に本稿で提案した手法を用いた詳細な分類体系を作成と，その体系の有効な利用法および評価をしていく．

参考文献

[1] P. Brown, V. Della Pietra, P, de Souza, j. Lai, and R, Mercer, Class-Based n-gram Models of Natural Language, Computational Linguistics 18 (1992), no.4, pp467 – 479.

[2] John D. Lafferty, Robert L. Mercer, Automatic Word Classification Using Features Of Spellings, Manuscript, IBM T. J. Watoson Reseach Center (1993).

[3] K. Church and P. Hanks, Word associatioon norms, mutual information, and lexicography, Computational Linguistics 16 (1990), no. 1.