先行詞探索による文章内容の分類

近藤恵子　東京工芸大学
上里福美　東京工芸大学

1.概要

文脈理解には、辞書、意味解析など多くの問題がある。この解決の一つの糸口として、文章をその内容により分類することを考えた。分類により辞書の範囲をしぼることが可能となり、また意味解析についても複数の選択肢に対する有効な条件となりうる。分類に当たっては、その使用されている名詞に着目した。文章中、重要な名詞は意味的には繰り返し使用される。しかし、実際文章を書くに当たっては繰り返し表現を避けるため、指示詞に変えられていることが多い。[1]では、文章中の指示詞を先行詞に置換する手法を提案した。今回はこの手法を用い、文章中の指示詞を先行詞に置換した後、文章に含まれる一般名詞の使用頻度を調査する。これにより、文章における各名詞の重要度を量る。辞書には予め、その名詞の属する分類内容、例えば「芸術」「数学」「経済」などの分類情報を記録しておく。各名詞の重要度と、辞書に記された分類内容から、文章の内容がどの分類に最も近いかを検討する。

2.先行詞探索手法

2.1.概要

文脈理解の手法構築にあたっては、指示詞の照応する先行詞を如何に検出するかが重要な課題となる。[2]では先行詞照応条件について幾つかの仮説を示し、検証を試みているが、具体的な検討は行っていない。そこで、その仮説に基づいた先行詞照応手法の構築と精度について検討した。

2.2.仮説

[2]では、照応について、以下の4つの仮説があげられている。

(仮説1)照応は段落にまたがることがない。
(仮説2)照応は交差することがない。
(仮説3)複文および重文の先頭を除く節中の照応に対する先行詞は、その節および、その節の上位の節には含まれない。
(仮説4)単文中の照応、および、重文の先頭節で、しかも埋め込み節でない節中の照応に対する先行詞は、それより前の文（段落をまたがってもよい）にある。

2.3.探索

探索は、以下の規則に即して行った。

（1） 複文、重文を単文に直す。
（2） 単文中の指示詞の場合は、前文より探索をする。
（3） 重文の第1文の指示詞の場合は前文より探索する。
（4） 一文による複文の下位文の指示詞の場合は、前文より探索する。
（5） 重文の第2文以後の下位文の指示詞の場合は、重文の前の文より探索する。
（6） 探索中に、前の指示詞に行き当たった時は、その先行詞まで飛び、そこから前へと解析を進める。
（7） 段落の始めまで遡っても照応する先行詞がない時、探索は失敗する。

2.4.結果

仮説のみの探索規則によるシミュレーションの正答率は、約33％であった。これには複数の原因が考えられ、その問題点についての改良を行った。指示詞に前もって与えた単数、複数などの条件は優先させつつ、ある程度の自由度を持たせた。連体詞の指示詞の場合、被修飾語が同じもの、もしくは関連したものが有効であるという条件を付加した。また、倒置文を検出、一文として訂正する。以上の改良により、正答率は約83％にアップした。

3.使用回数の調査

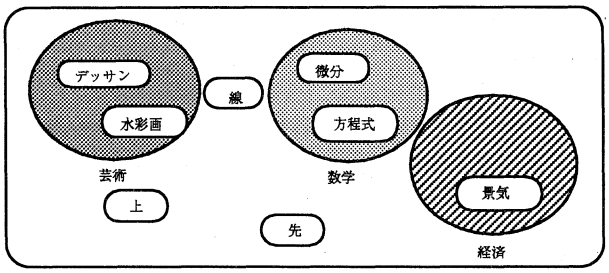
指示詞を先行詞に置き換えたことにより、語の面からの文意が明確になった。これは、指示詞によって意味的に繰り返し使用されていた語が、単語として明らかにされたことによる。この置換された後の文章における単語の使用頻度の高

さは、置換以前の文章中における各単語の意味的な重要度と連動する割合が高い。
今回の分類の手法はキーワードによる方法のため、調査は名詞のみを対象とした。手順を以下に示す。

（1）文章の頭より名詞を探索し、発見した名詞にはマーキングをする。
（2）文章の終わりに向かいながら、名詞を検索する。
（3）発見した名詞が過去にマーキングされていれば個数を数え、そうでなければ、また新たにマーキングを行う。
（4）これを、文章の終わりまで繰り返す。
（5）結果、文章中に使用されたすべての名詞に対して、使用回数が判明する。

4.分類方法

使用回数を利用し、分類を行う。分類は「芸術」「経済」「工業」「数学」等程度の範囲とした。
分類のキーワードとなる名詞には、辞書により分類情報を付した。キーワードとは明らかにその分野の専門用語である名詞を指す。例えば、「デッサン」という名詞には「芸術」の分類情報が付され、分類のキーワードとなる。「景気」という名詞には「経済」の分類が付され、キーワードとなる。この辞書のモデルは以下のように図示される。



名詞分類のモデル1

このキーワードの分類情報により、各分類の確率を計算する。計算は、以下の方法により行う。

$$\text{名詞Aの重要度} = \text{名詞Aの文章全体での使用回数} \qquad (1)$$

$$\text{「芸術」分類確率} = \frac{\Sigma\,(\text{「芸術」のキーワード名詞Aの重要度})}{\text{すべての分野に対するキーワードの出現回数}} \qquad (2)$$

この式は、その特定の文章中に現われたすべての分野のキーワードの使用回数のうち、「芸術」に関するキーワードの使用回数がどれほどであるかを表わしている。分子は文章中に使用された「芸術」に対するキーワードの出現回数と言い替えることもできる。他の分野に対しても、同様の計算でその確率を求めることができる。

文章名	総名詞数	キーワード使用回数		
		芸術	数学	経済
文章A	302	21	4	0

文章Aのキーワード使用回数調査の1例

この例においては、分類は以下の計算により行われる。これを全分野について計算し、最終的に最も確率の高いもの、この文章Aについては「芸術」に分類される。

「芸術」分類確率 ＝ 21／（21＋4＋0） ＝ 0.84
「数学」分類確率 ＝ 4／（21＋4＋0） ＝ 0.16
「経済」分類確率 ＝ 0／（21＋4＋0） ＝ 0

5.分類方法の改良

この方法ではキーワードの少ないものに対しては、精度が低くなる恐れがある。特に、専門語を避けて書かれた初心者向けの入門書について、その傾向が強い。そのため、専門性の多少低い名詞についても、分類に活用することを考えた。先の使用回数の調査において、キーワード以外の名詞に対してもその使用回数は明らかになる。この使用回数は、その文章に対するその名詞の重要度を表わしているという点では、キーワードも他の名詞も同様であると言える。しかし、

キーワードと他の名詞との明確な違いは、キーワードはある特定の分類にのみ含まれ、それ以外の名詞はいくつもの分野にわたり使用されるという点である。そのため使用頻度が如何に高くとも、分類に利用するに際してはキーワードと同等の重要度は持っていないと言える。よって（1）式は成立しない。
ここで、キーワード以外のその名詞が、分類に際してどれほど重要となるか、その重要度に重みを加重することを考えた。重みは学習により与え、式としては次のように書かれる。

名詞Aの分類αに対する重要度 ＝
名詞Aの文章全体での使用回数 × 名詞Aの分類αに対する重み （3）

5.1 重みの学習

重みは、キーワード以外のある特定の名詞Aの使用頻度が、分野によってどれだけ片寄っているかにより示す。「線」というキーワードでない名詞を例に考える。まず、各種の分野の文章について、先の分類方法により使用回数を調査し、計算、比較、分類を行う。この分類結果のモデルを以下に示す。ここでは特に「線」という名詞についてのみを扱った。

分類	文章名	総名詞数	「線」使用回数	文類別総名詞数	文類別「線」使用回数	文類別「線」使用頻度
芸術	文章A	302	32	458	39	0.085
	文章C	156	7			
数学	文章B	123	13	233	16	0.069
	文章E	110	3			
経済	文章D	108	0	108	0	0

分類結果のモデルの1例

文類別「線」の使用頻度とは、各分野での総名詞数に対する「線」の使用回数の割合を表わしている。「線」の「芸術」に対する重みは「線」使用頻度の全分野の合計のうち、「芸術」における使用頻度が占める割合により求められる。

「線」の「芸術」に対する重み ＝ 0.085／(0.085＋0.069) ＝ 0.55

同様の計算により、「数学」に対する重み〈0.45〉、「経済」に対する重み〈0〉を得る。この学習結果は新たな入力文章に対して働き、次の入力文章Fに「線」という名詞が含まれていた時、その分類の計算は、以下のように行われる。

文章名	総名詞数	キーワード使用回数			「線」使用回数
		芸術	数学	経済	
文章F	101	34	2	0	13

文章Fの使用頻度調査

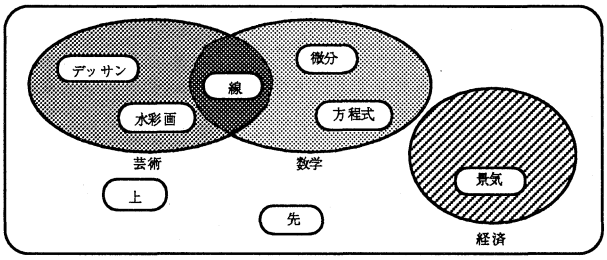
$$「芸術」分類確率 = \frac{34+(13\times 0.55)}{34+2+0+13} = 0.84$$

$$「数学」分類確率 = \frac{2+(13\times 0.45)}{34+2+0+13} = 0.16$$

よって、文章Fは「芸術」に分類される。
この結果よりさらに重みの学習をさせる。上の『分類結果のモデルの1例』の表にさらにこの文章Fを加えることにより、分類「芸術」の総名詞数は〈559〉、「線」使用回数は〈52〉、よって使用頻度は〈0.093〉となり、重みも以下のように変化する。

「線」の「芸術」に対する重み ＝ 0.093／(0.093＋0.069) ＝ 0.57

これにより他の分野に対する重みも変化し、「数学」に対する重みは〈0.43〉となる。
このときに使用した辞書のモデルは以下のように図示される。一つの分野の円の中に収まっている「水彩画」「方程式」などの名詞はキーワードであり、複数の分野の重なりにある「線」のような名詞は重みを加重することにより、二次的なキーワードとなりうる。

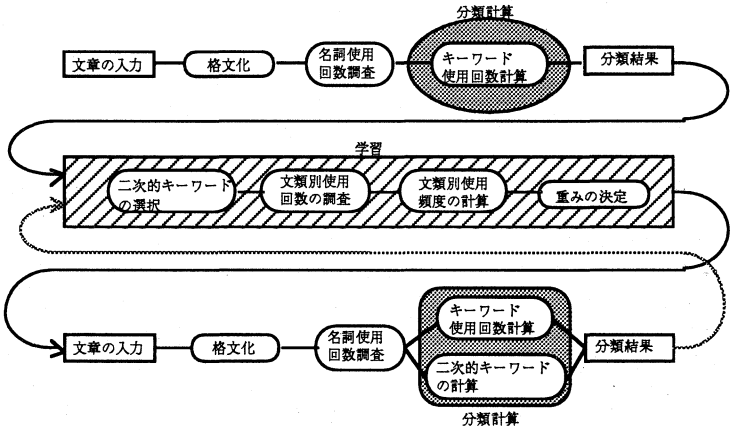


名詞分類のモデル2

このような二次的キーワードは複数の分野にまたがり使用されるが、ある程度の片寄りのあるものが望ましい。そのため、選択に当たっては各分野にわたる多くの文章の分類を行った後、蓄積された名詞の使用頻度の情報を元に、片寄りの大きい名詞を選ぶ必要がある。

6.構成

以上の手順を流れ図により示す。



手順の流れ模式図

7.まとめ

文章の分類に当たっては、キーワードの使用頻度が手がかりとなりうる。また、それ以外の名詞に対しても、重みを付加することにより、二次的なキーワードとなりうるものがある。この重みは各分野にわたる学習を繰り返すことにより、より精度の高い設定が可能となり、分類をより確実なものとすることができる。

参考文献

（2）「先行詞探索手法」
近藤恵子、上里福美　電子情報通信学会 1994年秋季大会講演論文集 情報・システム　D-65,p68,1994
（1）「日本語文章における照応・省略現象の基本的検討」
藤沢伸二、増山繁、内藤昭三　情報処理学会論文誌　Vol.34　No.9,1993