ラフな意味情報に基づいた文章の自動分類

金　明哲・宮本　加奈子
札幌学院大学

1. はじめに

文章(書)の分類に関する研究は大きく文章を書き手毎に分類することと文章を内容毎とに分類することに分けられる。前者は書き手不明の文章について書き手を判別・推定する研究である。これらの研究に関しては、金(1998, 1999)により追跡すればよい。後者は電子化された文章の検索や文章をテーマ・内容毎に自動分類することに関する研究であり、河合(1992)、徳永・岩山(1994)、湯浅・他(1995)、福本・他(1996)、新谷・他(1997)、福本・他(1998), Yamamoto, K.(1995)等がある。

内容による文章の分類・検索に関する研究はキーワードを用いる文章の分類・検索、全文を用いた文章の分類・検索に分けられる。キーワードによる方法は文章の分類・検索は速いが全文を用いるよりよい結果が得られることはかなり困難である。全文の情報を用いる方法としては、(1)文章に表れる単語の統計情報のみによる方法、(2)シソーラスなどの辞書を用いる方法が考えられる。(1)に関しては湯浅・他 (1995)は単語の共起に関する情報を用いて分類を試み、分類の正解率が 60%台であると報告されている。福本ら(1996)は辞書の語義文を用いて文書のクラスタリングを試み、湯浅らの単語の共起に関する情報を用いた分類より高い正確率(72.5%)を得た。このような文章の自動分類の報告結果は満足な正確率が得られたと言いがたい。文章の中の単語をシソーラスなど辞書に対応させる方法で文章の分類を行う方法は決して新しい提案ではないが、実際の研究報告が少なくい。シソーラス辞書を用いる方法に関して、河合(1992)はシソーラス辞書を用いる方法に関して、「分類体系への割り当てを人手で行うため効率が悪く、汎用性に欠けている」と指摘し、Niwa, Yら(1995)はシソーラスのカテゴリー自身が抽象的な語彙で定義されているため、文章の種類によっては、カテゴリー自身が文章に出現しない場合があると指摘している。上記のような指摘はシソーラスの辞書に関する問題点であり、着実に解決していかなければならないことである。このようなことが原因であるためか不明であるが、現時点での自然言語処理のソフト及び電子化辞書を用いた日本語の文章の自動分類に関する試みの報告は見られない。そこで、本研究では、現時点での形態素解析ソフト、電子化されたシソーラス辞書を用いたラフな意味情報に基づいく文章の自動分類を試みた。その結果、ラフな意味情報であっても、分類の技法を工夫することにより高い正確率で文章を自動分類することが可能であるという結果を得た。

2. 研究の材料と手順

2. 1 用いた材料

異なる１１人（本学の４年生）に下記の１０テーマについて、それぞれの文章の長さは約１０００文字以上になるように書かされた作文をテストの材料とした。そのテーマを下記する。０．住まい、１．家族、２．友達、３．学校、４．スポーツ、５．車、６．旅行、７．アルバイト、８．映画は映画館でみるか、ビデオでみるか、９．日本の食。

2. 2 データの加工

文章の形態素解析は JUMAN(黒橋・長尾 1998)を用いた。文章の機械的な自動分類を目指すため、形態素解析結果については一切手を入れなかった。

表１に用いた各文章の単語数を示す。

表１：分析に用いた作文ごとの単語数

著者／ﾃｰﾏ	0住まい	1家族	2友達	3学校	4ｽﾎﾟｰﾂ	5旅行	6車	7ｱﾙﾊﾞｲﾄ	8映画は…	9日本の食
A	605	665	937	610	680	614	610	745	636	596
B	565	617	552	652	741	672	655	609	667	578
C	630	925	623	735	559	550	658	707	676	586
D	658	565	658	585	573	644	610	593	610	569
E	556	591	711	636	545	641	609	565	579	624
F	583	607	540	558	549	572	603	632	551	629
G	607	696	810	704	622	597	664	705	556	672
H	571	615	672	645	661	561	716	545	642	686
I	585	566	577	623	586	575	567	571	599	605
J	566	578	586	566	596	580	609	599	554	617
K	767	580	601	550	568	571	568	585	659	587

文章を自動分類する際には、文章の中からどのような情報を抽出して用いるかが分類結果を大きく左右する。書き手による分類の場合はかなり複雑であるが、内容・テーマ毎の分類の場合は名詞がもっとも有効であると考えられる。本研究では名詞以外に動詞についても分析を試みた。

2. 3 方法

同一の話題・内容について異なる筆者が文を作成する場合、用いる単語が必ずしも同じではないことは誰でも理解に苦しまない。そこで、本研究では文章の中に現れる単語の完全一致をマッチするのではなく、シソーラス辞書を用いて同じカテゴリーに属する単語は内容において同質のものと考えることにした。例え、シソーラス辞書のあるカテゴリー $C_i(w_1, w_2, \cdots w_z)$ は z 個の単語により構成されているとする。文章中の単語 $w_j \in C_i$、$w_g \in C_i$ である場合、w_j と w_g は内容におけるラフな同質のものであると判断する。

シソーラス辞書としては、国立国語研究所の「分類語彙表」を用いた。分類語彙表の階層情報は用いなかった。表２に分類語彙表の演劇・映画のカテゴリーを示す。

表２「分類語彙表」のしくみ

1.3240 劇　軽演劇
1. トラマ　メロドラマ　ホームドラマ　芝居　村芝居　人形芝居　現代劇　時代劇　コミック　ﾊﾟﾗｴﾃｨｰ　等
2. 歌劇　野外劇　屋内劇　舞台劇　ﾊﾟﾄﾏｲﾑ等
3. 雅楽　田楽　猿楽　能　歌舞伎　ﾐｭｰｼﾞｶﾙ　ﾚﾋﾞｭｰ等
4. 映画　劇映画　洋画　邦画　キネマ　シネマ　ｻｲﾚﾝﾄ　文化映画　記録映画　ﾄﾞｷｭﾒﾝﾀﾘｰ　ｱﾆﾒｰｼｮﾝ　動画　八ミリ　本編　リメーク等

文章中の単語とシソーラス辞書との対応によ

るデータの作成は具体的には下記ように行った。

(1) それぞれの文章に現れるすべての単語（名詞、動詞）を品詞毎に分け分類語彙表のカテゴリーへ対応させ、カテゴリー単位で頻度を求める。つまり、文章中の単語(名詞、動詞に分け)w_iについて $if \quad w_i \in C_j \quad then \quad FC_j = FC_j + 1$とする。ここの$C_j$はシソーラスの辞書の$j$番目のカテゴリーであり、$FC_j$は分類語彙表の中のカテゴリー$C_j$中の単語が文章中に現れた頻度である。ただし、

a. 分類語彙表に登録されていない単語は「その他１」というカテゴリーを設け、「その他１」のカテゴリーに属すると判断する。
b. 表記が一致しないため機械的に対応するカテゴリーが存在しないと判断されている単語は「その他１」のカテゴリーに属すると判断する。
c. ある単語が分類語彙表の中の複数のカテゴリーに属する場合は、それぞれのカテゴリーに頻度をカウットする。
d. 一つの文章における集計が終わった時点で、変数の数を減らすため一つのカテゴリーの相対頻度が３以下の場合は、「その他２」のカテゴリーに併合させる。

このような基準に基づいた文章iにおけるカテゴリーに関する頻度ベクトルを下記のように表記する。

$$F_i = (FC_{i1}, FC_{i2}, \cdots FC_{ij} \cdots FC_{ig})$$

(2) 各文章の単語数は必ずしも同じではないので（1）で求められた頻度ベクトルを相対頻度ベクトルに置き換える。上記のカテゴリーによる頻度ベクトルから下記のような相対頻度ベクトルを求める。

$$P_i = (p_{i1}, p_{i2} \cdots p_{ij} \cdots p_{ig})$$

ここでは $p_{ij} = FC_{ij} \Big/ \sum_{j=1}^{g} FC_{ij}$ である。

(3) すべての文章における相対頻度ベクトルを用いて相対頻度データマトリックスを作成する。各文章で求められた相対頻度ベクトルのカテゴリー種類及び数は必ずしも一致しない。例え、i行のベクトルのj番目のカテゴリーがl行のベクトルには存在しない場合、l行のベクトルにj番目のカテゴリーを挿入し、そのデータはゼロとする。したがって、下記のマトリックスの中のnと相対頻度ベクトルのgとの関係は$n \geq g$である。

$$P_{m \times n} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix}$$

(4) データマトリックスから距離マトリックスを作成する。ここではK-L-S距離を用いた。i行のデータが、$p_{i1}, p_{i2} \cdots p_{ij} \cdots p_{in}$、$l$行のデータが、$p_{l1}, p_{l2} \cdots p_{lj} \cdots p_{ln}$のとき、$i$行と$l$行との間の距離$d_{il}$は下記のような式を用いて定義される。

$$d_{il} = \frac{1}{2} \sum_{k=1}^{n} \left(p_{ik} \log \frac{2 p_{ik}}{p_{ik} + p_{lk}} + p_{lk} \log \frac{2 p_{lk}}{p_{ik} + p_{lk}} \right)$$

$p_{ik} = 0$ ならば $\quad p_{ik} \log \dfrac{2 p_{ik}}{p_{ik} + p_{lk}} = 0$，

$p_{lk} = 0$ ならば $\quad p_{lk} \log \dfrac{2 p_{lk}}{p_{ik} + p_{lk}} = 0$

3. 実験分析

2章の手順により得られた相対頻度のマトリックス、距離のマトリックス用いて、主成分分析、多次元尺度法、階層的クラスターリング、判別分析などの分析を行い、学習情報がない場合、学習情報がある場合の文章の自動分類の正確率などについて実験分析を行った。(同日に発表する)

4. おわりに

本研究では、11人の１１０編の作文を用いて、ラフな意味情報による文章の自動分類を試みた。その結果、意味情報がラフであるのにも関わらず約９０％前後の正確率で文章を自動に分類することが可能であるという結果が得られた。本研究では主に

(1) 名詞は文章の内容に関する重要な情報となる。
(2) 動詞は名詞に比べ文章の内容に関する情報が少ない。
(3) ラフな意味情報だけでも高い正確率で文章を自動分類することが可能である。

等の知見が得られた。今後の課題としては、先行研究に提案された様々な方法と組み合わせることによる正確率のアップなどが挙げられる。

謝辞

JUMAN の使用に関しては、京都大学黒橋禎夫様に、「分類語彙表」については「分類語彙表」作成の代表国立国語研究所の中野　洋様にお礼を申し上げます。本研究は札幌学院大学社会情報学部理系プロジェクトの研究助成金を受けました。

参考文献

[1] 金　明哲(1998). 日本語における計量文体学の近年の進展、INFORMATION An International Journal, Vol.1, No.2, 57-64.
[2] 金　明哲(1999). 日本現代文における書き手の特徴情報、人文学と情報処理、No.30.
[3] 河合 敦夫(1992). 意味属性の学習結果に基づいく文章の自動分類について、情報処理学会論文誌、Vol.33, No.9, 1114-1122.
[4] 徳永 健伸・岩山 真(1994). 重み付き IDF を用いた文章の自動分類について、情報処理学会自然処理研究会、Vol.100, No.5, 33-40.
[5] 湯浅 夏樹・上田 徹・外川 文雄(1995). 大量の文章データ中の単語間共起を用いた文書分類、情報処理学会論文誌、Vol.36, No.8, 1819-1827.
[6] 福本 文代・鈴木 良弥・福本 淳一(1996)辞書の語義文を用いた文章の自動分類、情報処理学会論文誌、Vol.37, No.10, 1789-1799.
[7] 福本 文代・鈴木 良弥・福本 淳一(1998)意味的類似性と多義性解消を用いた文章検索手法、自然言語処理、Vol.4, No.3, 51-69.
[8] 新谷 研・角田 達彦・大石 巧・長尾 眞(1997) 単語の共起頻度と出現位置による新聞の関連記事の検索手法、情報処理,」Vol.38, No.4, 855-862.
[9] 湯浅 夏樹,外川 文雄: 概念識別子の頻度分布を利用した文書分類, 情報処理学会研究報告 Vol.95 No.87 (95-FI-39), pp.33-40, 1995.
[10] Niwa, Y. and Nitta, Y.(1995): Statistical word Sense Disambiguation Using Dictionary Definitions, Natural Language Processing Pacific Rim Symposium '95, Seoul, Korea, 665-670.
[11] 国立国語研究所(代表：中野　洋).『分類語彙表』東京: 電子版, 1996.
[12] Yamamoto, K.. Masuyama, S. and Naito, S.(1995). Automatic Text Classification Method with Simple Class-Weighting Approach, Proceedings of Natural Language Processing Pacific Rim Symposium '95,498-503.
[13] 黒橋 禎夫 長尾 真(1998): 日本語形態素解析システム JUMAN version 3.5, http://www.ish.ic.kanagawa-it.ac.jp/~okamoto/Documents/Juman/manual/manual.html.