共起情報を利用した新聞記事の自動分類結果の分析・評価

4Q－5

藤井洋一、今村誠、高山泰博、鈴木克志

三菱電機株式会社　情報技術総合研究所　音声・言語インタフェース技術部

1.はじめに

近年大量のテキスト情報がインターネットなどを通じてアクセス可能となるにつれて、蓄積された文書の分類整理を目的とする文書の自動ファイリングへの要求が高まっている。特に文書内容による自動ファイリングが望まれる。そのための技術としてベクトル空間モデルがある。

ベクトル空間モデルによる自動分類は自動学習可能なため、大量のテキストデータを扱う場合に適している。しかし、ベクトル空間モデルによる分類は単純に単語の出現頻度のみを捕らえて分類先を決定するため分類精度に問題があり、精度向上のための方式が提案されている。例えば、河合[1]は単語の意味属性を用いて分類精度が向上することを示し、福本[2]はクラスタリングにおいて辞書の語義文を用いた多義解消により精度が向上することを示した。

我々は、[1]で指摘されている単語の多義性を共起情報を用いて解消すれば精度向上すると予測した。ただし、従来の言語学で言われる「多義語」を多義解消するのではなく、分類が詳細になった場合顕著となる複数分類項目で頻繁に出現する単語に注目した(例えば、単語「大統領」は<首相>や<大統領選挙>といった複数の分類項目で頻繁に出現する)。この、複数分類項目で頻繁に出現する単語を「分類多義語」と定義し、共起情報を用いて「分類多義語」の多義を解消した自動分類を試みた[3][4]。

今回、本方式で自動分類した結果と、河合の方式のうち単語のみの情報で自動分類(従来方式と呼ぶ)した結果とを比較し、本方式が有効な場合を明らかにする。また、従来方式でも本方式でも解決できない点を分析する。

Analysis and Evaluation of the Automatic Articles Classification Using Lexical Co-occurrences
Youichi FUJII, Makoto IMAMURA,
Yasuhiro TAKAYAMA, Katsushi SUZUKI
Human Media Technology Dept. Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan

2.実験方式の概要と分類結果

2.1.従来方式

従来方式は、学習記事から分類先の単語出現頻度を学習し、χ^2統計を応用した方式で重み付けする。分類項目を$C_i(i=1,\ldots,N)$、単語をw_j $(j=1,\ldots,L)$、出現頻度をF_{ij}とすると、理論頻度を(1)、重み付けを(2)で計算する。

$$M_{ij}=\sum_{i=1}^{N}F_{ij}\cdot\sum_{j=1}^{L}F_{ij}\Big/\sum_{j=1}^{L}\left(\sum_{i=1}^{N}F_{ij}\right) \quad (1)$$

$$Y_{ij}=(F_{ij}-M_{ij})\cdot\left|F_{ij}-M_{ij}\right|\Big/M_{ij} \quad (2)$$

分類対象記事の単語出現頻度を$D=(d_1,\ldots,d_L)$とし、$S_j=(Y_{1j},\ldots,Y_{Nj})$とした時に、$\sum_{j=1}^{L}(S_j\cdot d_j)$を計算して$(s_1,\ldots,s_N)$とおく。$C_i$への類似度を$s_i/(\sum_{i=1}^{N}s_i)$として分類先を決定する。

2.2.本方式

上記、従来方式で学習した重み付けに対し、(3)を満たす単語$w_j(j=1,\ldots,L)$を分類多義語とする。

$$\#\{Y_{ij}\mid\max_{1\le k\le N}(Y_{kj})\cdot 0.1\le Y_{ij}\}\ge 2 \quad (\#\text{は要素数}) \quad (3)$$

次に(3)を満たす分類多義語を、図 1のように分類項目付き単語に分割し、(2)式で重みを再計算する。また、分類項目付き単語に付与された分類先の学習記事中から段落内共起する単語を抽出する。

学習テーブルの頻度学習結果

単語 / 分類項目	大統領
首相	120
大統領選挙	50
裁判	5
地方行政一般	70
外交関係	90
：	4

再学習テーブルの頻度学習結果

大統領 <首相>	大統領 <大統領選挙>	大統領 <地方行政一般>	大統領 <外交関係>
120	0	0	0
0	50	0	0
1.25	1.25	1.25	1.25
0	0	70	0
0	0	0	90
1	1	1	1

図 1　分類多義語の頻度分配例

自動分類時には、分類対象記事中の単語出現頻度を求めるとともに、分類多義語に対しては分類対象記事中から同様に段落内共起単語を取り出す。学習時の共起情報と分類対象記事の共起情報間で内積値を計算し、内積値が大きいものだけで分類対象記事の分類多義語の出現頻度を比例分配した後、従来方式と同様に類似度計算して分類先を決定する。

2.3.実験結果

「朝日新聞」記事(朝日新聞社提供)[5]１年分をランダム抽出により学習記事61500記事と分類対象記事3857記事とに分割し、734分類項目への自動分類実験をした。形態素解析には、JUMAN 2.0[6]+EDR日本語単語辞書1.5版[7]を利用した。

新聞記事には複数の分類項目が付与されているので、１記事の出現頻度は分類先の数で割って各分類先に学習させた。名詞および、固有名詞、未知語、サ変名詞のみ使用した。実験結果を表 1に示す。

表 1　新聞記事１年分の分類結果

	精度(%)
従来方式	48.1
本方式	51.4

精度は再現率と適合率が等しい点(分類数 10948)

3.3%の精度向上(正しい分類先が366増加)した。

3.分析・評価

3.1.正解／不正解数の変化から見た分析・評価

上記分類結果を正解数の変化と不正解数の変化を軸にして分類項目毎に分類したのが図 2である。

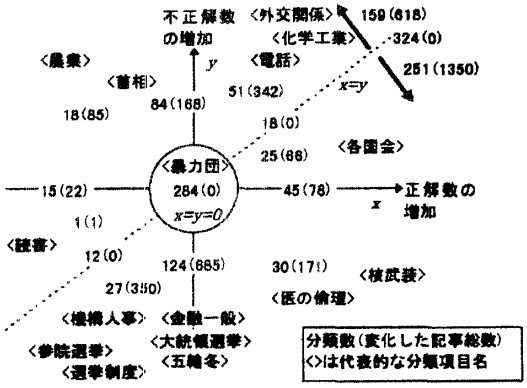

図 2　分類結果の変化の分布

直線$x=y$より下側の251分類が精度向上に寄与し、159分類が精度低下に影響し、324分類は精度に影響しなかった。また、本方式は不正解数を減少させる方向に働くことで分類精度向上に寄与していることが変化した記事総数から分かる。実際、<大統領選挙>は従来方式では単語「大統領」のもつ分類多義性が影響し156記事が誤分類された。本方式によって正解記事数は変わらず、誤分類が109記事減った。

一方、<暴力団>は単語「暴力団」が<暴力団>のみで多く現れるため分類多義語とならず、<暴力団>に誤分類した記事を本方式でも修正できなかった。

3.2.多義語から見た分析・評価

言語学で言われる多義語として単語「ホーム」を考える。分類多義語解消の効果として、このような単語も多義解消される可能性がある。分類対象記事中 16 記事で「ホーム」が出現した。多義解消の結果を表 2に示す。概ね正しく多義解消されている。

表 2　多義解消結果の例

分類＼記事中の意味	鉄道のホーム(8)	施設としてのホーム(5)	その他(3)
その他企業,ＨＡ	1	3	1
国鉄,私鉄,鉄道事故	20	1	1
福祉一般,老人,児童福祉	4	10	6

単位は頻度分割で割り振られた回数　括弧内は記事数
（１記事に対して複数分類への割り振りあり）

3.3.分類項目の性質から見た分析・評価

新聞記事の分類項目名には、明確な分類先が存在しない場合に与える<○○一般>という分類が存在する。再現率と適合率が等しい時点での分類結果を、<○○一般>に限定して評価した結果を表 3に示す。

表 3　<○○一般>の分類精度

	再現率(%)	適合率(%)
従来方式	34.5	37.4
本方式	38.4	40.3

これらの分類項目は単語の出現傾向の特徴量が弱く、分類精度が上がらなかったと考えられる。

4.まとめ

分析・評価の結果、分類項目が多い場合でも本方式が誤分類を減少させる方向で分類精度の向上に寄与している事が分かった。しかし、本方式は複数の分類項目に高頻度で現れる単語を処理することしかできないため、「暴力団」のように特定の分類項目にのみ現れた単語によって誤分類した結果は修正できない。また、[2]で提案されているような同義語の統一を行っていない。

上記の課題を解決する方式が必要である。

[参考文献]
[1] 河合:意味属性の学習結果に基づく文書自動分類方式，情報処理学会論文誌，Vol.33，No.9(1992).
[2] 福本他:辞書の語義文を用いた文書の自動分類，情報処理学会論文誌，Vol.37，No.10(1996).
[3] 藤井他:共起情報を利用した文書の自動分類について，情報処理学会第54回全国大会，3分冊，4K-10(1997).
[4] 藤井他:共起情報を利用した文書の自動分類，情報処理学会自然言語処理研究会，NL118-6(1997).
[5] 朝日新聞記事データベース(1991年9月～1992年8月).
[6] 松本他：日本語形態素解析システムJUMAN使用説明書 version 2.0(1994).
[7] EDR電子化辞書 日本語単語辞書1.5版，（株）日本電子化辞書研究所(1995).