出現情報を用いた新聞記事ジャンルの推定システム

藤本 太郎　　菅野 道夫
東京工業大学 総合理工学研究科 システム科学専攻
tarof@fz.dis.titech.ac.jp

1 はじめに

今日の通信インフラの発達などによって、個人でも大量のデータを入手可能になったが、各個人の処理能力に限界がある。そこで、本稿では個人が情報処理を行なう前段階での支援システムとして、新聞記事 [1] 情報をジャンル毎に分類を行なうシステムについて説明を行なう。

まず、本研究では機能言語学 [4] の立場より、ジャンルを「使用目的別の分類」と定義付け [2]、アンケートによるジャンル分類を行なう。次に、入力された新聞記事テクストのジャンル推定の方法として、従来方法として名詞ベクトルによる方法 [3] を用いた方法と、本研究で提案する出現頻度を用いた方法および両者を組み合わせた方法の3種類の手法について評価を行う。状況情報がジャンル推定に与える効果と従来手法とファジィ合成を行なうことによる推定精度向上について議論する。

1.1 ジャンルとレジスター

選択体系機能文法 [4] ではジャンルを「ある特定の文化において達成する目的別の分類」と定義しており、それぞれのジャンルはジャンルを特徴付ける固有のレジスターを持つと考えられている [2]。

また、レジスターとは談話のコンテクスト [4] の
・状況のフィールド (何が話されているか？)
・テナー (誰と誰が話しているか？)
・モード (どのような媒体を使用しているか？)
の3項組である。

1.2 ファジィ演算

ファジィ頻度表現: コーパスから得られる頻度情報を反映する一手法である。平均以下の値を切り捨てる方法と出現確率値の折衷であり、式1で表現される [5]。

$$Freq_x = \begin{cases} 1 & (If \ \ x \geq \bar{x}) \\ \frac{x}{\bar{x}} & (If \ \ x < \bar{x}) \end{cases} \quad (1)$$

max・min 合成演算: 各ジャンル毎に最小の類似度 (min) を取り[1]、その中でもっとも大きい値 (max) のジャンルを選ぶ方法である [6]。

2 ジャンルの定義

1.1節の定義に基づいて、学生10人に対して新聞記事におけるジャンルを挙げてもらった。その結果、以下に挙げる18種類のジャンルを得た。比較のため、他手法におけるジャンルを表1に挙げる。

国内政治	国際政治	中央経済	地方経済
国際経済	中央社会	地方社会	国際社会
スポーツ	コラム	企業情報	科学技術
医療	商品情報	テレビ	論説・解説
人物・インタビュー		文化・芸術・娯楽	

3 提案するジャンル推定システム

与えられた新聞記事が、アンケートによって選られた18種類のジャンルのどれに当てはまるかを以下の3種類の手法を用いて推定を行う。
・名詞ベクトルを用いる手法 [3]
・出現情報 (出現紙面) を用いる方法
・名詞ベクトルと出現情報を用いる方法

[1] 今回の場合は名詞ベクトル法による類似度と出現情報法による類似度

	分類数	ジャンルのラベル
湯浅 95	4	政治、経済、事件、国際
福本 96	8	産業・経済、科学技術・文化、金属・土石、流通・サービス、小売業、環境・公害、化学、農林・水産
Kessler 97	6	レポート、論説、法律、科学技術、ノンフィクション、小説
本手法	18	国際経済、中央政治、地方社会ほか

表 1: 他手法におけるジャンルとの比較

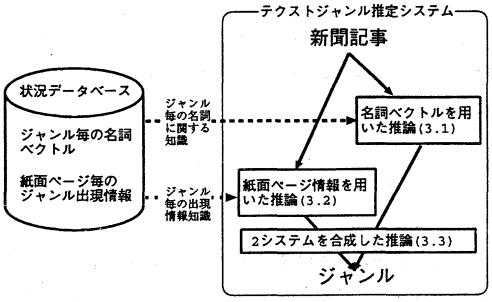

図 1: テクストジャンルシステムの構成

3.1 名詞ベクトルを用いたジャンル推定システム

学習:学習用データから、高頻度の 1024 名詞を抽出し、各ジャンル毎に 1024 名詞の出現数を求め、分野ベクトルとする。

推定:入力された記事の記事ベクトル（1024 次元）を求め、類似度として記事ベクトルと分野ベクトルの正規化した内積を取り、もっとも類似度が高いジャンルをその記事のジャンルとする。また、記事毎にジャンルとの類似度をファジィ頻度表現し、値が 1 となるものをその記事のジャンルにする。

評価:表 2および表 5に示すように、湯浅の結果 [3] と比較して適合率が低くなっている。これは分類数が 4 から 18 に増えたためと考えられる。また、ファジィ頻度表現によって再現率が向上し、適合率が低下している。これはファジィ頻度表現によって類似度が平均以上のものがジャンルとして選ばれ、複数解答を許しているためと考えられる。

	再現率	適合率
名詞ベクトル法	30.4%	30.4%
ファジィ頻度表現	87.5%	13.1%

表 2: 名詞ベクトル法を用いた場合の評価

3.2 出現情報を用いたジャンル推定システム

学習:まず、学習用データから紙面ページとジャンルの対を作成する。次に、出現頻度表現データとして各紙面毎にジャンルの出現頻度を求める。さらにファジィ頻度表現データとして出現頻度表現をファジィ頻度表現する。

推定:まず、入力記事の紙面ページ情報を抽出する。次に、出現頻度表現データとマッチさせて最も高い出現頻度のものをその記事が属するジャンルとする（出現頻度法）。また、ファジィ頻度表現データとマッチさせて最も高いファジィ頻度のものをその記事が属するジャンルとする（ファジィ頻度表現）。

評価:表 2と表 3を比較すると、出現情報を用いる方法は名詞ベクトル法と比較して高精度である。このことにより、出現情報はジャンル推定に有効な推定キューであると言える。しかし、ファジィ頻度表現された出現情報の再現率が低いため、ファジィなどを用いた欠落データ補完 [7] の必要があると考えられる。

	再現率	適合率
出現頻度法	39.2%	37.2%
ファジィ頻度表現	67.2%	18.7%

表 3: 出現頻度を用いた場合の評価

3.3 二種類の推定システムを結合させた推定システム

合成: 名詞ベクトル法と出現情報をファジィ頻度の使用・不使用によって4通りの組み合わせを作り、max・min 合成演算して類似度を計算する。

・名詞ベクトル法+出現情報
・名詞ベクトル法のファジィ頻度表現+出現情報
・名詞ベクトル法+ファジィ出現情報
・名詞ベクトル法のファジィ頻度表現+ファジィ出現情報

評価:表4に示すように、適合率が最も高いものは名詞ベクトル法+ファジィ出現情報を用いた場合である。つまり、従来手法である名詞ベクトルを用いた方法 [3] に対し、出現情報をファジィな制約として加えることにより、精度が向上したといえる。出現情報をファジィ頻度表現することによって緩やかな制約となり、名詞ベクトル法と協調して推論を行うことができるといえる。

	再現率	適合率
名詞ベクトル+出現情報	32.4%	28.5%
名詞ベクトル (fuzzy)+出現情報	39.3%	37.2%
名詞ベクトル+出現情報 (fuzzy)	48.2%	48.2%
両方 fuzzy	66.1%	28.9%

表 4: 2システムを結合させたシステムの評価

3.4 他手法との比較

表5に示すように、本手法はもっとも適合率が低いが、これは分類数が多いためと考えられる。そのため、分類数を6に減らすと、適合率は66.1%となり湯浅 [3] や、テクストタイプを用いる Kessler[8] の手法より精度は高まる。名詞関係辞書を構築する福本 [9] の手法には劣るが、構築する知識ベースが単純という利点を有する。

4 結論

本研究では、従来手法である名詞ベクトル法 [3] にファジィ頻度表現を施した出現情報を加えたテクストジャンル推定システムを構築した。出現情報を加えることによって、適合率が30.4 %から48.2 %に向上し、推定精度が向上することを示した。

	分類数	適合率
湯浅 95	4	63.8%
福本 96	8	72.5%
Kessler 97	6	64.7%
本手法	18	48.2%
	6	66.1%

表 5: 評価の比較

今後の課題として、名詞ベクトル (フィールド情報) に関して、福本 [9] のように名詞関係辞書を構築したり、高頻度名詞ではなくジャンルを特徴づけやすい名詞を選ぶことによっても精度向上が期待できる。また、出現情報 (モード情報) に関して学習用データに欠けている出現情報を補完する機構 [7] を構築することで推定精度を向上させられると考えられる。

参考文献

[1] 日本経済新聞社. 日本経済新聞 *CD-ROM* 版. 日本経済新聞社, Mar. 1995.

[2] 塚田 浩恭. 体系機能言語からみた日本語文法. PhD thesis, 関西外国語大学大学院外国語学研究科言語文化専攻, 1996.

[3] 湯浅 夏樹 上田 徹 外川 文雄. 大量文書データ中の単語間共起を利用した文書分類. 情報処理学会論文誌, Vol.36 No.8, Aug. 1995.

[4] M. A. K. Halliday. *AN INTRODUCTION TO FUNCTIONAL GRAMMAR.* Edward Arnold, London, second edition, 1994.

[5] 藤本 太郎 and 菅野 道夫. 名詞概念との共起関係を用いた用言概念の分類. 言語処理学会 第 *3* 回年次大会予稿集, Mar. 1997.

[6] 大里 有生 本多 中二. ファジィ工学入門. 海文堂, Jul. 1989.

[7] K Hirota L.T. Koczy. Ordering,distance and closeness of fuzzy sets. *Fuzzy Sets and Systems*, 59, 1993.

[8] B. Kessler, G. Nunberg, and H. Schutze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of ACL*, pages 32–38, 1997.

[9] 福本 文代, 鈴木 良弥, and 福本 淳一. 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, 37(10):1789–1799, 1996.