E-013

ニュース分類を考慮した自動タギング手法

A Tagging Method Based on News Classification

オウ　ヨンヒ† Yeonhee Oh　佐野 雅規ǂ Masanori Sano　藤井 真人ǂ Mahito Fujii　柴田 正啓ǂ Masahiro Shibata

1. はじめに

ウェブ 2.0 の登場で、誰もが容易に自分の文章をインターネット上に公開できるようになった。最近のこれらの文章には、タグ（tag）という一種のメタデータが人手で付与され、検索用のインデックスをはじめ、内容分類や関連性の判断など、様々な用途に利用されている。本稿では、ニュース記事を対象として、このタグを自動付与する手法を提案する。一般的なタグは、単独の名詞だけでなく、複数の形態素から形成される句も含んでおり、これらは内容を理解するのに有用とされている。そこで本稿では、ルールによる名詞や名詞型句の生成と、ニュース分類による固有名詞の選択を組み合わせたタグ自動付与手法を提案する。また、本手法により生成したタグに関する主観的評価の結果について述べる。

2. 提案手法

図 1 にニュース記事からのタグ抽出手法を示す。図に示すように、2 つの抽出方法を並列に用いる。1 つは、まとまった意味を持つ名詞や名詞型句の表現の抽出、もう 1 つは、ニュース分類によりあらかじめ設定したルールに基づいて固有名詞を抽出するものである。以下に、各処理の詳細を説明する。

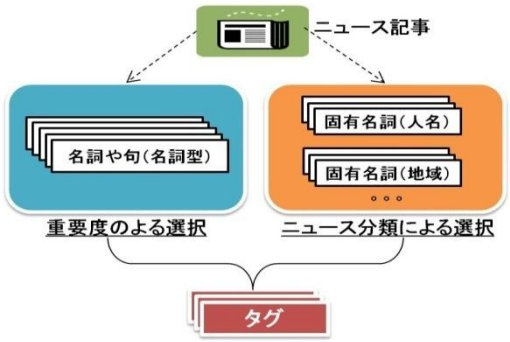

図 1 ニュースタギング提案手法

2.1. 意味のある名詞や名詞型句の抽出

ニュース記事の各文章から名詞や名詞型句を抽出するために、cabocha[2]の形態素分析結果を利用し、以下のルールで抽出する。

（1）名詞や名詞型句の始め
- 名詞、形容詞-自立、接頭詞

例外）名詞-接尾-助動詞語幹、名詞-副詞可能、名詞-代名詞-一般、名詞-非自立、名詞-ナイ形容詞語幹

（1-1）名詞や名詞型句の始め（次の品詞が名詞のとき）
- 名詞-形容動詞語幹

（2）名詞や名詞型句の続き
- 名詞、助詞-連体化（‘の’のみ）、助詞-格助詞-一般（‘や’のみ）、接頭詞、形容詞-自立

例外）名詞-非自立-副詞可能

（2-1）名詞や名詞型句の続き（次の品詞が名詞のとき）
- 記号-一般（‘・’のみ）、助動詞（‘な’のみ）、助詞-格助詞-一般（‘と’のみ）

（3）名詞や名詞型句の終り（次の品詞前）
- その他（動詞、助詞、名詞-接尾-助動詞語幹、名詞-副詞可能、名詞-代名詞-一般など）

抽出された名詞や名詞型句の重要度(PI)は次のように計算する。句に含まれる単語の TF-IDF の値を加算し、更にニュース内の句の出現頻度数を掛け合わせた値である。これは句を形成する各単語の重要度と句の頻度の両方が反映された値であると考えられる。

$$PI(p, d, D) = |\{p_k : p_k = p\}| \times \left(\sum_{i=1}^{n} TFIDF(w_i, d, D)\right) \quad ①$$

- $|\{p_k : p_k = p\}|$：ドキュメント d で句 p の出現頻度数
- w_i：句 p が n 個の単語を含む場合の i 番目の単語
- $TFIDF(w, d, D) = |\{t_j : t_j = w\}| \times \log \frac{|D|}{|\{d : w \in d\}|}$
- $|D|$：ニュースの総ドキュメント数

2.2. ニュース分類による固有表現抽出

ニュース記事のタグとしては、単独の固有名詞も重要であると考え、利用することとした。固有名詞には、いくつもの種類が存在し、またニュースの内容によりこれらの重要度が違うと考えた。そこで、まずニュース記事の分類を判定し、あらかじめ設定した分類ごとのルールを用いていくつかの固有名詞をタグとして選択する。以下に、分類判別の手法と、設定したルールについて説明する。

2.2.1. ニュース分類の自動判別

一般的に、ニュース分類には曖昧な部分があるが、放送局や新聞社は記事の内容や出所などいくつかの基準によりニュースを分類している。本稿での分類は表 1 のように定義する。

表 1　ニュース分類

ニュース分類	主要内容
政治	政府、政党、政治活動、選挙など
経済	経済指標、企業、農水産業など
社会	事故、事件、災害など
国際	日本と関係ない海外ニュース
スポーツ	野球、サッカー、ゴルフなど

各ニュース記事を表 1 のように分類するために、SVM(Support Vector Machine)を利用する。用いる特徴は、ニュース記事の中の名詞であって、かつ該当する分類にだけよく出現する名詞を使う。しかし、該当する分類におい

†韓国放送 放送技術研究所　KBS BTRI

ǂNHK 放送技術研究所　NHK STRL

てその名詞の出現するニュースの数が極端に少ない場合には、その名詞は使用しない。この条件を数式で表したものが式②である。ニュース記事をベクトルで表現する単語 w_i の選択基準②と特徴ベクトル③の数式は次のとおりである。

$$\frac{WA(w_i, D(x))}{WA(w_i, D(All))} > \alpha \quad and \quad WA(w_i, D(x)) > \beta \quad -②$$

• $WA(w_i, D(x)) = \frac{|\{d:w_i \in d, D(x)\}|}{|D(x)|}$, WA: 分類 x での単語 w_i を含むニュースの比率、$|D(x)|$：分類 x の総ドキュメント数（例. x : 政治)

$$V(d) = \{Val(w_1, d), Val(w_2, d), \ldots, Val(w_N, d)\} \quad -③$$

• $Val(w, d) = \frac{|\{w_j : w_j = w\}|}{Length\ (d)}$, Val: ドキュメント d で単語 w の特徴量

2.2.2. ニュース分類による固有名詞選択ルール

ニュース記事においては、人名や地域などの固有名詞は有用な情報であり、タグの有力候補である。しかし、ニュース分類により、これら各固有名詞の重要度には違いがあると考えられる。本研究ではこの重要度を表すものとして、各分類の中に現れる主な固有名詞の頻度数の平均を用いることとした。各分類において約 60 のニュース項目を用いて、主な固有名詞の頻度数の平均を求めたものを表２に示す。あるニュースから固有名詞をタグとして抽出する場合、そのニュースの分類の平均頻度の数に従ってタグを選ぶ。

表 2　ニュース分類による固有名詞選択ルール

	人名	機関	国名	地域
政治	2	4	1	1
経済	0	2	1	1
社会	1	1	0	4
国際	1	1	2	1
スポーツ	2	1	1	1

3.　実験結果

実験に用いた番組は NHK ニュース７である。ただし、IDF の計算には、実験データを含む 1 年間のニュースの単語コーパスを用いた。

3.1.　句表現の抽出結果

100 のニュース項目から抽出した句表現の数を、TF-IDF により抽出した単語数と比較した。ニュース記事の平均文字数は 526 であり、平均 TF-IDF 単語数は 54、平均句表現数は 45 であった。句表現数が 45 になった理由は、2 個以上の単語が組み合わされて 1 つの句になるという減少要因と、1 つの単語が前後の形容詞などと組み合わされて複数の句が抽出されるという増加要因の総合的な結果である。

3.2.　ニュース自動分類結果

SVM ツールは libsvm[3]を利用した。SVM の訓練データは 5 つの分類に対して約 60 個(毎月 5 項目づつ抽出)のニュース項目を使用した。数式②の α と β は、それぞれ‘2.0’と‘0.03’を使ったときに結果が一番良かったので、以下の実験ではこれらの値を用いた。テストデータとしては 2 週間分のニュースを用いた。実験結果を表３に示す。

表 3　ニュース分類成功率

	政治	経済	社会	国際	スポーツ
成功率	91.4 %	89.7 %	85.1 %	84 %	96.6 %
統合	70 %（SVM 出力の最も高い分類を選んだ場合）				

3.3.　総合評価

タギング結果を評価するため 100 のニュース項目を使用し評価者による実験を行った。実験は各ニュース項目から本研究によって自動付与したタグと、TF-IDF の値の上位のものから自動付与したタグの数と同数の単語を選んで比較評価するものである。タグは句抽出の結果から 5 個と、ニュース分類により表２の該当する数だけ選んだ固有名詞を使用した。比較のため抽出された TF-IDF 単語とタグの例を以下に示す。

比較評価の基準は「どちらがニュース記事の内容をよく表現しているか」とした。実験には 6 人が参加した。表４に結果を示す。実験に参加した人のほとんどが、本手法によるタグのほうが良いと答えた。

TF-IDF 単語：1) ５年ぶり, 2)マイナス, 3)株, 4)株価, 5)逆風, 6)値下がり, 7)年初比
タグ：1) ５年ぶり, 2)サブプライムローン問題, 3)株価, 4)東京, 5)日本, 6)日本株, 7)年初比マイナス

表 4　各評価者による結果

	人１	人２	人３	人４	人５	人６
TF-IDF 単語	20	47	27	15	16	6
タグ	79	42	67	85	81	93
評価困難	1	11	6	0	3	1

この結果の信頼性をさらに確認するために、各ニュース項目について 6 人の評価の平均を計算した。その結果を表５に示す。この結果を見ると 6 人はほぼ同様の判断をしたものと考えられる。

表 5　各ニュースに評価を総合した結果

	TF-IDF 単語	タグ	等しいなど
ニュース数	9	80	11

4.　おわりに

本研究では、ニュースのタグを自動付与するためにニュース分類による固有名詞タギングと名詞や名詞型句抽出によるタギングを組み合せた技術を提案した。この技術による検索結果は、今後、ニュースのクラスタリングに活用していく予定である。同時に、より多くの評価者の意見をもとにタグの品質を向上させることも必要であると考えている。

今回の実験では、固有名詞と一般的なタグを 5 つ選択した。しかし、いくつのタグが必要かを決めることは難しい問題である。実験に参加した人にアンケート調査した結果、1 つのニュース項目には 5〜10 くらいのタグが適切だとの回答が多かった。また、人名などの固有名詞の場合は、各分類別に 1〜3 が有用だとの答えが多かった。一方で、「名前は有名人の場合だけ必要であり、それ以外の人は性別や年齢の情報で十分」との意見もあった。

参考文献

[1] G. Salton and M. J. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983
[2] 工藤拓、松本裕治, “チャンキングの段階適用による日本語係り受け解析”, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834－1842, 2002
[3] C. Chang and C. Lin, “A library for support vector machines”, 2001