構文情報を用いたテキストの自動分類

鈴木　祐介*，松原　茂樹，吉川　正俊 (名古屋大学)

Text Classification based on Lexical Dependencies

Yusuke Suzuki, Shigeki Matsubara, Masatoshi Yoshikawa (Nagoya University)

1　はじめに

インターネット上に存在する大量の Web 文書があらかじめ内容別に整理されていれば，分野や内容を絞った検索が可能になる．従来のテキスト分類手法の多くは，単語の出現頻度を用いて分類カテゴリやテキストを特徴付けてきた [1]．本稿では，構文情報を用いたテキスト分類手法を提案する．文の依存構造に着目し，単語間と文節間の依存関係を特徴素に加えることにより，構文的特徴を考慮したテキスト分類が実現できる．

2　構文情報を用いたテキスト分類

本手法では，従来の単語による特徴素以外に，依存関係にある単語の対も特徴素として抽出し，これら 2 種類の特徴素を用いてテキストを分類する．依存関係としては文節内の各単語の依存関係を表す単語間依存関係と，文節間の依存関係を表す文節間依存関係を用いる．

2.1　依存関係にある単語対の抽出

依存関係にある単語対の抽出例を Fig.1 に示す．単語間依存関係の抽出では，日本語の文節内の単語は後方修飾が原則であり，かつ，大半が直後の単語を修飾するという観察に基づき，文節内で隣り合う 2 つの単語を単語対として抽出する．文節間依存関係の抽出では，係り受け解析器によって文節間の依存関係を求め，依存関係にある文節内の主辞に相当する単語を単語対として抽出する．

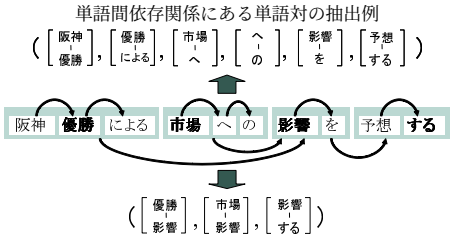

Fig. 1: Word-pairs which form dependency relations

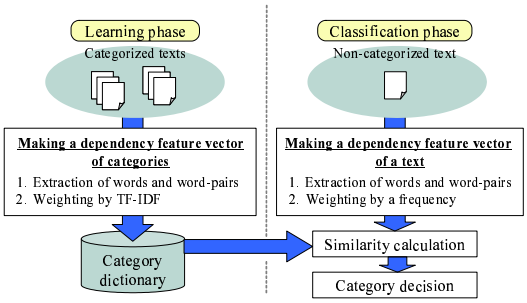

Fig. 2: Flow of text classification

2.2　分類方法

テキスト分類は，各カテゴリの特徴ベクトルを作成する学習フェーズと未分類文書をカテゴリに分類する分類フェーズからなる．テキスト分類の流れを Fig.2 に示す．

学習フェーズでは，カテゴリ C_i における単語 e_j(j=1,2,$\cdots$,N) の TF-IDF による重み w_{ij} と，単語対 p_j(j = 1,2,$\cdots$,L) の TF-IDF による重み d_{ij} を式 (1) で定義し，カテゴリ C_i における単語と単語対を特徴素とした特徴ベクトル $(w_{i1}, w_{i2}, \cdots, w_{iN}, d_{i1}, d_{i2}, \cdots, d_{iL})$ を作成する．

$$w_{ij} = \frac{F^w_{ij}}{\sum_{j=1}^{N} F^w_{ij}} log \frac{M}{v^w_j} \qquad d_{ij} = \frac{F^d_{ij}}{\sum_{j=1}^{L} F^d_{ij}} log \frac{M}{v^d_j} \tag{1}$$

ここで，F^w_{ij} はカテゴリ C_i における単語 e_j の出現頻度，v^w_j は単語 e_j を含むテキスト数，F^d_{ij} はカテゴリ C_i における単語対 p_j の出現頻度，v^d_j は単語対 p_j を含むテキスト数，M は学習データの総数である．

分類フェーズでは，テキスト t における単語 e_j の出現頻度による重み w_{tj} と，単語対 p_j の出現頻度による重み d_{tj} を式 (2) で定義し，テキスト t の単語と単語対を特徴素とした特徴ベクトル $(w_{t1}, w_{t2}, \cdots, w_{tN}, d_{t1}, d_{t2}, \cdots, d_{tL})$ を作成する．

$$w_{tj} = F^w_{tj} \qquad d_{tj} = F^d_{tj} \tag{2}$$

ここで，F^w_{tj} はテキスト t における単語 e_j の出現頻度，F^d_{tj} はテキスト t における単語対 p_j の出現頻度である．テキスト t のカテゴリ C_i における類似度 $sim(t, C_i)$ を式 (3) で定義し，値が最大となるカテゴリをテキストの分類先カテゴリとする．なお，α はパラメタである．

$$sim(t, C_i) = \sum_{j=1}^{N} w_{ij} w_{tj} + \alpha \sum_{j=1}^{L} d_{ij} d_{tj} \tag{3}$$

3　評価実験

実験では，分類カテゴリに Yahoo!のカテゴリ (14 カテゴリ) を使用した．また，各カテゴリに登録されているサイトのトップページ (HTML 文書) を収集し，学習データ 902 ページ，テストデータ 99 ページに振り分けた．実験では，形態素解析器として Chasen[2] を，係り受け解析器として Cabocha[3] を使用している．また，式 (3) の α は 1.2 とした．

その結果，本手法で 51.5%の正解率を示した．特徴素に単語のみを用いる従来手法の正解率 48.5%と比べて 3.0%上昇しており，テキスト分類に構文情報を用いる効果を確認した．

4　まとめ

本稿では，構文情報を用いたテキスト分類手法を提案した．実験により，テキスト分類に構文情報を取り入れることの有効性を確認した．

文　献

(1) 森本．他：情報処理学会研究報告，DD6-1，pp.1-8，1997.
(2) 松本．他：NAIST-IS-TR9908, 1999.
(3) 工藤, 松本：情報処理学会論文誌, 43(6), pp.1834-1842, 2002.