潜在意味に基づく新聞記事からの特徴抽出

奥村, 直也 / OKUMURA, Naoya

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
58
(開始ページ / Start Page)
1
(終了ページ / End Page)
2
(発行年 / Year)
2017-03-31
(URL)
https://doi.org/10.15002/00014244

潜在意味に基づく新聞記事からの特徴抽出に関する研究

STUDIES ON FEATURE EXTRACTION FROM NEWS ARTICLES BASED ON LATENT SEMANTICS

奥村 直也
Naoya OKUMURA

指導教員 三浦孝夫

法政大学大学院理工学研究科システム理工学専攻修士課程

In recent years, we can get huge documents from the internet. These documents have huge amount of information. Very often, in news articles, headlines contain characteristic expressions specific to their contents. In this investigation, we propose new methods to generate headlines for news articles with automatic estimations. Here we examine both news articles and the headlines separetely, bridge two documents in such a way that similar articles have similar headlines. In experiments, we show the effectivities of our approaches by some experimental results.
Key Words *: News Articles, Latent Semantics, Headline Generation, Dependency Structure, Feature Extraction*

1. 問題の背景

近年，インターネットで電子書籍や新聞記事が容易に入手できる．短期間のうちにあまりに大量な文書が生じるため，何ら利用されることなく流れ去っている．これらの文書は，多様で大量のデータ量を含み，内容やその傾向を理解することは困難であり時間を要する．このため，文書内容を素早く理解するために，文書のラベル付けや文書要約などの特徴抽出手法が必要となる．

文書のラベル付けでは，キーワードとなる単語の集合をラベルとして抽出を行う．ラベルは，文書の内容の特徴部を表す表現である．ラベルの役割は，複数文書の場合では「どのジャンルに分類されているか」の判断材料になり，単文書の場合では「大まかにどういった内容を含んでいるか」の判断材料になる．文書が大規模である場合，人手による文書のラベル付けはコストが大きいため，自動的にラベル付けされることが望ましい．しかし，自動ラベル付けでは，キーワードとなる単語の意味を考慮していないため,「正しい解釈がされているか」という問題がある．

文書要約では，類似する文書中から適切な情報を抽出し，得られた情報から要約文章の抽出や生成を行う．要約は，記事本文の特徴を表現したものであり，文書の素早い理解に有効である．利用者は，正しい要約を読むことで，記事本文を読まず，1 記事の全貌を把握することができる．このため，要約により文書の内容を明確化することで，分類や検索が容易かつ効率的に行える．文書集合が大規模である場合，人手による文書要約はコストが大きいため，自動的に文書を要約できることが望ましい．

類似文書の検索では，文書中の単語の出現頻度 TF (Term Frequency) やその拡張の TF*IDF (Inverse Document Frequency) となる重みを付け，余弦類似度などにより検索を行う．しかし，この検索手法では，文書の意味を考慮していないため,「適切な文書を選択できているか判断できない」という問題がある．

要約文章の抽出または生成では，得られた情報から，文章のランキング付けなどを行うことで要約文章の抽出を行ったり，文章を文法に基づく手法を利用することで 1 から要約文章の生成を行う．しかし，文書から要約として，自動的に文章を 1 から生成することは，文章の意味を考慮しなければならないため，困難となっている．

本研究では，文書の特徴抽出手法として，単語の意味を考慮した自動ラベル付けや，自動要約を行う．

2. 扱う問題

(1) ニュース記事クラスタの知的ラベル付け

文書のラベル付けでは，キーワードとなる単語の集合をラベルとして抽出を行う．文書集合が正しくラベル付けされれば，内容の素早い理解や検索が可能となる．キーワード抽出では，特徴的な単語の抽出に出現頻度や共起頻度，エントロピーに基づく手法が研究されている．しかし，これらの手法は，単語の意味である語義や語同士の関連性を扱うことが難しく，特徴的な表現を表すことが難しい．

本研究では，ラベル語集合を上位概念で統一し，文書集合の意味を考慮したラベル付けを提案する．これにより，単語の意味を考慮したラベル付けが，適切なジャンルを選択できていることを示す．[1] [2] [4]

(2) 潜在意味解析を用いた新聞記事の見出し生成

文書の要約は，文章の生成や抽出によって行われる．文章を選ぶ際，他の類似する文書から得られる情報を利用して，文章の生成や抽出を行う研究されている．類似文書の検索では，文書中の単語に出現頻度などの重みを付け，余弦類似度により検索を行う．しかし，この検索手法では,「文書の意味を考慮していない」という問題がある．

本研究では，潜在意味解析を利用し，類似した文書本文を選択し，その文書の見出しを用いることで新たな半自動見出し生成を提案する．これにより，文書の意味を考慮した類

似文書の検索が，適切に見出しを選択できていることを示す．[3] [5]

(3) 係り受け関係を用いた新聞記事の見出し生成

文書の自動要約は，要約を行いたい文書から，生成された文章や抽出された文章を要約とする．要約文章の生成では，ビタビアルゴリズムを用いることで，1 から要約文章の生成が研究されている．要約文章の抽出では，ページランクを用いたランキングベースによる要約文章の抽出を行ったり,, Support Vector Regression (SVR) を用いた学習ベースによる要約文章の抽出が研究されている．文書から要約として，自動的に文章を 1 から生成することは，構文解析などの文法解析を行い，文章の意味を考慮しなければならないため，困難である．

本研究では，特徴的表現を考慮するために，類似した見出しのパターンを使用することで，その対応を考察する．潜在意味解析を利用することにより，類似した文書を選択し，対応する見出しに対して係り受け関係を使用し，キーワードを入れ替えることにより自動見出し生成を行う．これにより，係り受け関係を利用することで，適切な見出しを生成できることを示す．[6]

3. 結果

まず，ニュース記事クラスタの知的ラベル付けでは,「スポーツ」タグが付与された記事 1433 件を「どのスポーツについての単語集合がラベル付けされるか」のラベル推定を行う．正解率は提案手法は 35 %，ベースラインは 8 %となっている．提案手法が考慮できていない点は，ベースラインでは具体的などのスポーツか判断できる単語が選択されているが，提案手法では一般的な単語が多いため，どのスポーツか判断できないものが存在する．

次に，潜在意味解析を用いた新聞記事の見出し生成では，記事 1761 件の見出し推定と特徴語抽出を行う．見出し推定では，正解率は提案手法は 92 %， ベースラインは 72 %となっている．特徴語抽出では，正解率は提案手法は 57 %，ベースラインは 40 %となっている．提案手法が考慮できていない点は，見出し推定では，提案手法は社説記事中に特徴的なパターンがほとんど現れないため，類似記事を検索できない場合が存在する．特徴語抽出では，特定の話題の記事が 1 記事しか存在せず，特徴語がうまくとれない場合が存在する．

最後に，係り受け関係を用いた新聞記事の見出し生成では，記事 5055 件の実験 1：特徴語抽出と実験 2：段落による特徴語抽出の二つの実験を行う． 実験 1：特徴語抽出では，提案手法とベースラインを比較評価するために，記事に合った特徴語が抽出できたかと特徴語として正しいかを評価する．図 1 より，正解率は提案手法 46 は %，ベースラインは 39 %となっている．実験 2：段落による特徴語抽出では，第 1 段落のみの記事が検索範囲として有効かを確かめるために，第 1 段落と第 2 段落を含めた記事を使用し，記事に合った特徴語を抽出できたかを評価する．正解率は第 1 段落のみでは 46 %，第 1 段落と第 2 段落では 39 %となっている．提案手法が考慮できていない点は，特徴語抽出では，係り受け関係で抽出できない単語がベースラインよりも多くなる場合が存在する．段落による特徴語抽出では，見出しの特徴語を入れ替える箇所が，少なかったため，選ばれる見出しが悪い場合が存在する．

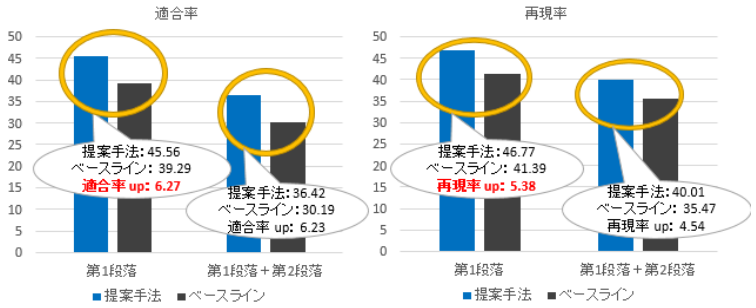

図 1 結果

4. 結論

本研究では，文書内容を素早く理解するために，文書のラベル付けや文書要約を用いることで，文書の特徴抽出が行えることを示した．特徴抽出では，文書の内容をおおまかに知りたいのか，細かく知りたいのかにより，必要な手法が異なる．

ニュース記事クラスタの知的ラベル付けでは，単語の意味を考慮したラベル付けが，適切なジャンルを選択できていることを示した．潜在意味解析を用いた新聞記事の見出し生成では，文書の意味を考慮した類似文書の検索が，適切に見出しを選択できていることを示した．係り受け関係を用いた新聞記事の見出し生成では，係り受け関係を利用することで，適切な見出しを生成できることを示した．

参考文献

1) Okumura, N. and Miura,T.，”Automatic Labelling of Documents Based on Ontology”，IEEE Pacific Rim Conference on Communications，Computers and Signal Processing (PACRIM) 2015，Victoria，Canada
2) Okumura, N. and Miura,T.，”Labelling Document Using Ontology”，5th Intn’l Conf. on Digital Information Processing and Communications (ICDIPC) 2015，Sierre/Siders，Switzerland
3) Okumura, N. and Miura,T.，”Headline Candidates for News Articles”，5th IEEE International Workshop on Data Integration and Mining (DIM-2016) 2016，Pittsburgh，USA
4) 奥村　直也，三浦　孝夫，”ニュース記事クラスタの知的ラベル付け”，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM），福島，平成 27 年（2015） 3 月
5) 奥村　直也，三浦　孝夫，”潜在意味解析を用いた新聞記事の見出し生成”，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM），福岡，平成 28 年（2016） 3 月
6) 奥村　直也，白井 匡人，三浦 孝夫，” 係り受け関係を用いた新聞記事の見出し生成”，第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM)，高山，平成 29 年 (2017) 3 月