平成 27 年度 公立はこだて未来大学卒業論文

トピックモデルによる単語の属する話題の推定手法

田中 桂介

情報アーキテクチャ学科　　1012047

指導教員　　新美 礼彦

提出日 平成 28 年 1 月 29 日

Word Topic Prediction Based on a Topic Model

by

Keisuke Tanaka

BA Thesis at Future University Hakodate, 2016

Advisor: Ayahiko Niimi

Department of Media Architecture
Future University Hakodate
January 29, 2016

Abstract− In this study, a prediction method for word topics is proposed. Sentences are sometimes difficult to read if the meaning of one of more words is unknown. In such cases, sentences become easily readable if the topics of the words are known. To solve this problem, a topic model that divides the words by topic and chooses several words from the available topics is used in this study to predict topics. This study suggests that the information regarding the chosen words can be used as a topic indicator. Several experiments were conducted using the proposed method with regard to the model, and the following two facts were established: the proposed method which divides the words by topic, shows an accuracy of ˜60% and the method that chooses several symbolic words for topics shows an accuracy of ˜90%.

Keywords: Text mining, Topic model, Latent Dirichlet Allocation, Document Classification

概 要:

本研究は, 文章中の単語の属する話題を推定する手法の提案を目的とする. 文章の読解において, 単語の示す意味を知らず, 文章の意味が理解できないことがある. そのような場合, 単語の厳密な定義を即座に知ることは難しくても, 単語が何に関する語であるのかを知ることができれば, 大まかに文章の意味をつかむことができる. そこで本研究では, データの背景に潜在している話題によってデータを分類することが可能なトピックモデルを単語の分類に取り入れた手法を提案した. 提案手法では, (1) 文章に形態素解析を行い, (2)LDA による分類モデルを作成して単語を属している話題ごとに分類し, (3) 分類された単語群の中から代表語を選出する, という手順となっている. これによって, 単語がどの話題に関するものであるのかを推定する. 提案手法の LDA による分類モデルの作成, 及び同一トピック内における代表語の選出に対するそれぞれの評価実験によって得られた結果から, 提案手法は多数の文章から 6 割程の精度で単語を背景に持つ話題ごとの正しい分類に成功すること, 背景に持つ話題ごとに分類された単語群から代表語を 9 割程の精度で選出できることがわかった.

キーワード: テキストマイニング, トピックモデル, LDA, 文書分類

目 次

第1章　序論

本章では, 本研究の背景, 目標, 論文の構成を述べる.

1.1　背景

私達は現在, 書籍や新聞, Web ページなどの媒体を通して, 多くの日本語の文章を読む機会を有している. そして文章の読解中に, 文章の示す意味が理解できず, 文章読解が困難になる問題が生じる.

文章の読解が困難になる問題の中でも, 本研究では, 文章を構成する単語の中に知らない新語や専門用語が含まれているなど, いずれかの単語が示す意味を知らない, 理解できないことに起因している場合に焦点を当てる. この場合には, 知らない単語の辞書的な意味を調べることで問題を解決することが理想となるので, 理解できない単語の定義を逐一調べていくことによって問題の解決を図れるが, それには辞書や単語の解説をしている Web ページなど, リソースとなる情報が用意されていることが前提となるので, そのような情報が手元に無い場合はこの方法による解決が期待できない. 加えて, 文章の意味が理解できるまで必要な単語の定義や解説を逐一全て, もしくは文章の理解に必要となる部分を探して読んでいくことには時間を要する.

また, 本研究で着目する点として, 単語の辞書的な定義を知ることはできなくても, 単語と同じ話題の文章で用いられる関連語, 単語の上位概念となる語など, 単語が属している話題の情報を大まかな意味や背景として知ることができれば, 文章全体が示す意味も大まかにつかむ事ができる点が挙げられる.

1.2　研究目標

背景で述べた, 文章を構成する単語の示す意味がわからず, それに連なって文章の示す意味の理解が困難になる問題を補助する方法のひとつとして, 辞書や単語の解説をしている Web ページの情報など, 単語の辞書的な意味に頼らずに単語が属している話題の情報となるような他の単語を選出することを目標とする.

そこで本研究では, トピックモデルに基づいて文書データから文章中の単語をトピックとして話題ごとに分類し, 分類した単語群の中から単語の重みを参照して代表となる単語を選び, 単語がどの話題に関するものであるのかを推定することで, 文章中の単語がどの話題に属しているかを示す手法を提案する.

また, 本研究でのトピックは, 単語が属する話題ごとに分類された単語のグループを示し, 本研究での単語の重みは, TF-IDF による重み付けから得られた単語に対する重みの値を示す.

1.3　論文の構成

この節では, 本論文の次章以降の構成について記述する. 第 2 章では, 本研究での根幹となる技術であるテキストマイニングやトピックモデルを中心とする手法, 及びその実装に利用するツールについて説明する. 第 3 章では本研究と関連する研究について説明する. 第 4 章では本研究で提案する手法について説明する. 第 5 章では提案手法についての評価実験とその結果について説明する. 第 5 章では本研究のまとめや今後の課題についてを述べる.

第2章　関連手法とツール

本章では, 本研究の根幹となる技術や手法, 及び本研究で用いるツールの説明を行う.

2.1　テキストマイニング

テキストデータに対する処理技術であるテキストマイニングに関する説明を行う.

2.1.1　テキストマイニング

大規模なテキストデータに対して処理を行うことで, 個々のテキストからでは得られない新たな情報や知識を得る技術であり, データマイニングをテキストデータに対して適用させたものがテキストマイニングである. テキストマイニングの中には, テキストデータに対して自然言語解析の手法を使って単語や文節で区切り, それらが出現する頻度や他の出現との相関や傾向などを解析することで, 文章をカテゴリごとに分類する分析方法がある [1].

2.1.2　TF-IDF

単語の出現回数である TF（Term Frequency）と一般語へのフィルタとして機能する IDF（Inverse Document Frequency）を掛け合わせた重みの一種が TF-IDF である. 単語がテキスト内で何回出現したかというベクトルが TF であり, これのみでテキストデータの特徴ベクトルとすることもできるが, TF のみを特徴量として考えると, テキストデータ内で話題を特徴づけるような重要な単語とどの文書にも出現する一般的な単語が同等に評価されてしまう. そこで, TF に IDF という単語を含むテキストデータが多いほど小さい値になるベクトルを掛け合わせて重み付けを行い, これをテキストデータの特徴量とすることにより, 多くのテキストデータ内で使用される一般的な単語は重要ではない単語とみなされ, より少ないテキストデータ内で使用される単語がそのテキストデータを特徴づける単語であるとみなす.

本研究では, テキストデータに含まれる単語に対してこの TF-IDF による重み付けを行った単語ベクトルを作り, そのベクトルを提案手法における単語の話題ごとの分類や単語群の代表となる単語選出の判断材料とした.

2.1.3 形態素解析

文書データに対して, 文章を形態素区切りで分割し, 文章を構成している単語の情報を得る処理が形態素解析である. 形態素解析を行うことで, 文章を文節や単語単位で出現頻度を集計することが可能になるので, その後のデータ処理や分析が行いやすくなる.

本研究での形態素解析は, 形態素解析ソフト MeCab[2] 及び R 言語の RMeCab パッケージ [3] を用いて行い, TF-IDF による重みの値を計算する.

2.1.4 コサイン類似度

ベクトル空間モデルにおいて, 文書ベクトル同士を比較して文書の類似度を計算する手法がコサイン類似度である. コサイン類似度の計算では 2 つの文書ベクトルの内積を計算してベクトル同士の成す角度の近さを表現するので, コサイン類似度が 1 に近いほど 2 つの文書が類似していることになる.

本研究では, 単語ベクトル同士を対象にコサイン類似度の計算を行い, 特定の単語に類似している単語を選出する.

2.2 トピックモデル

本研究で利用する確率的生成モデルであるトピックモデルに関する説明を行う.

2.2.1 トピックモデル

文書データの解析手法として提案された, 確率的生成モデルがトピックモデル (Topic model) である [4]. トピックモデルでは, データの集合にはその背景にあらかじめ隠れた話題や分野が存在していて, データはそれに従って分布されている, かつ 1 つのデータは複数の話題を併せ持っていると仮定して, そのうえで話題を推定し, データがそれぞれの話題に対してその話題に属している確率を求めることで, データの背景にある隠れた話題や, データがどの話題に属しているのかを推定していく [6].

トピックモデルでは, 文書データを出現する単語の順序関係を無視した頻度分布である BoW(Bag of Words) と呼ばれる多重集合で表現していて, その生成過程をモデル化している. これにより, 単語の並びに関する情報より文章中でどのような単語が使われているかを重視しながら文章の持つ話題を推定していく. また, この BoW 表現における文章と単語の関係を他のデータ形式に適用させることで, 画像処理、Web 解析といった他の分野への応用が可能である.

本研究では, トピックモデルによる, 文章にどのような単語が含まれているかの情報から文章の持つ話題を推定する工程を, 単語がどのような文章に含まれているかの情報から単語の持つ話題を推定するよう適用する.

2.2.2 LDA

トピックモデルに階層ベイズモデルを導入して, 一般化させたモデルが LDA(Latent Dirichlet Allocation) である [5]. トピックモデルの研究では, LDA の学習アルゴリズムに関する研究, LDA のモデルを拡張させる研究, LDA のモデルを応用させる研究が中心となっていて, 本研究はこのうち LDA のモデルを応用させる研究にあたる.

2.3 ツール

本研究で提案手法の実装に用いたツールに関する説明を行う.

2.3.1 R 言語

オープンソースで開発された, 統計解析向けのプログラミング言語及びその開発実行環境が R 言語 [7] である. R 言語では, ベクトル処理と呼ばれる実行機構により, ベクトルからデータフレーム, 時系列といった構造のデータを宣言無く変数に納められ, 処理を簡単に記述することができる. 統計に適した解析環境に加えてパッケージが充実しており, 導入することによって様々な統計処理や機械学習を行う関数を利用することが可能となる.

本研究での提案手法は, 形態素解析及び TF-IDF による重み付けは RMeCab パッケージ [3], LDA による分類モデルの作成及び分類モデルを利用した単語の話題による分類は MASS パッケージ [8], 単語同士のコサイン類似度の計算は proxy パッケージ [9] を利用して, R 言語によって実装した.

第3章　関連研究

本章では, 本研究と関連した研究として, LDA を文書の分類に応用させている研究, 単語の持つ印象を推定する研究について説明を行う. 加えて, 関連研究と本研究との違いについても説明を行う.

3.1　k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応

新納らの研究に, 自然言語処理のタスクにおいてある領域の訓練データから学習された分類器を, 別の領域のテストデータに合うようチューニングする, 領域適応を行う研究がある [10]. この研究では, 単語間の類似度が測れる仕組みを用意して単語のクラスタリング結果に対応させるための手段として, トピックモデルの 1 つである Latent Dirichlet Allocation (LDA) を利用して, 単語のソフトクラスタリングを行っている.

3.2　トピックモデルに基づく文書ストリームのマルチラベル分類

白井らの研究に, 文書ストリーム中の文書のラベルの特徴を動的に学習して, ラベル間の相関関係をラベリングに利用することで文書ストリームのマルチラベル分類を行う研究がある [11]. この研究では, トピックモデルを拡張させたモデルを提案し, 文書の持つラベルベクトルと語集合から, 単語を生成する潜在変数であるラベルとトピックを推定することで文書集合のトピック分布を学習させている. また, 未知の文書に対するラベルベクトルの推定を, 各ラベルから文書が生成される尤度から単一ラベルを求め, その単一ラベルと共起するマルチラベルのセットから尤度の高いマルチラベルを選択することで実現させている.

3.3　ヘルプデスク作業効率化のためのラベリング自動化

堀内らの研究に, Apple サポートコミュニティに投稿された質問文書に対して Wikipedia の記事タイトルを用いたラベル付けを自動で行う研究がある [12]. この研究では, Apple サポートコミュニティへの質問文書に LDA を適用させて得られた各文書のトピックの混合比とトピック毎の単語生成確率を, 別のコーパスである Wikipedia の記事集合に当てはめて適用させ, トピックの生成確率からラベルとなる記事タイトルを選択することによって, ラベリングの自動化を実現させている.

3.4　形容詞共起を用いた単語の印象推定法

清水らの研究に, 形容詞や形容動詞が持つ印象を単語同士の共起頻度から推定する研究がある [13]. この研究では, 形容詞, 形容動詞の共立共起に限定させた共起頻度の測定から類似度を求め, 類似度から印象に応じた数量化を行って空間上に単語を配置して表現することで, 単語の印象の推定を実現させている.

3.5　関連研究との違い

本研究の特色は, トピックモデルを単語単位に着目して適用させる点と, トピックの代表語を選出することでトピックへのラベリングを行う点にある.

関連研究 3.1 から 3.3 まではトピックモデルを文書に適用させて問題の解決を図っている研究である. 関連研究 3.1 では, 単語の類似度測定のための手段として LDA が用いられているが, 本研究では単語そのものの分類結果を得ることが必要となるので, LDA による処理過程が異なる. 3.2 では, トピックに対するラベリングに関して未知の文書に対してマルチラベルの推定を行っているが, 本研究では単一のラベルをトピック内の単語から選定する点が異なる. 3.3 では, ラベルの候補を別のコーパスのデータにして, その中から選定を行っているが, 本研究では文書データ中で用いられている別の単語から単語の属する話題を表現したいため, 同一コーパスのデータからラベルを選定する点が異なる.

関連研究 3.4 とは単語の持つ大まかな意味を推定する研究として目標が類似しているが, 本研究では推定の過程にトピックモデルを用いることによって, 単語の共起を用いた手法において考えられる, 単語の出現位置に推定結果が左右されてしまう問題への改善が期待できる.

第4章　提案手法

本章では, 本研究で提案する手法について説明を行う.

4.1　提案手法概容

本研究の提案手法では, 訓練用となる文書データを用意して, 以下の処理を行うことで全文章中の単語をトピックごとに分類し, 同一トピックに含まれる単語全てに対して単語がどの話題に属しているかを示す情報として, 分類されたそれぞれのトピックから代表語となる単語の選出を行う.

1. 訓練用の文書データに対して形態素解析を行い, 文章を構成している単語の情報を得る
2. 取得した情報から LDA による分類モデルを作成し, 文章中の単語をトピックごとに分類する
3. 手法の推定結果として, 分類したトピックから, 代表語を選出する

本手法は, 文書分類におけるトピックモデルを単語単位に着目して適用させる点や, トピック内に含まれる単語から代表語を選出することでトピックへのラベリングを行う点が特徴であり, 話題の推定過程にトピックモデルを用いることによって, 単語同士の前後関係や出現位置に左右されずに推定結果を出力できる点が利点である.

次節以降では, 提案手法の各工程についてのより詳しい説明を行う.

4.2　形態素解析

あらかじめ文章の持つ話題の情報が紐付けられている文書データに対して形態素解析を行い, 文章を構成している単語やその重み, 及び単語を含む文章が持つ話題の情報を得る. 形態素解析を行った後に, それぞれの単語に対して, その単語を含む文章が持つ話題の情報を参照し, 単語と話題の情報を紐付ける. 単語が複数の文章で用いられている場合は, 重みの値が最大となる文章が持つ話題の情報を紐付ける. 形態素解析は, 形態素解析ソフト MeCab[2] 及び R 言語の RMeCab パッケージ [3] を用いて行い, TF-IDF による重みの値を計算する.

4.3 LDA による分類モデルの作成

形態素解析によって得られた単語の情報から, LDA に基づいて単語からその単語を含む文章が持つ話題を推定する分類モデルを作成することで, 文書データ中に含まれる単語や未知の単語に対して, 単語をトピックごとに分類できるようにする. LDA の実装には R 言語の MASS パッケージ [8] を用いる.

4.4 代表語の選出

分類モデルにより分類されたトピックから, トピックのラベルとなる, トピックを代表する語を選出する. 選出された代表語をトピック内の単語における単語の属する話題の推定結果として出力し, これによって文章中の単語の示す意味から連なって文章の示す意味の理解が困難になる問題の解決へのアプローチを図る. 代表語の選出については, TF-IDF による単語の各文章における重みの合計値が高い 3 つの単語, 及びその補助として重みから選出された 3 つの単語それぞれに対するコサイン類似度が高い 3 つの単語を重複を許して選出し, 選出された 12 個の単語を推定結果として出力する. コサイン類似度の計算には R 言語の proxy パッケージ [9] を用いる.

第5章　実験と評価

本章では, 本研究で行った提案手法の LDA による分類モデルの作成工程に対する性能の評価実験 (実験 1), 代表語の選出工程に対する性能の評価実験 (実験 2), 及び実験 2 の評価基準の正しさを検証する補助実験 (実験 3) の結果, 及び考察を述べる.

5.1　実験概容

提案手法の評価実験に関しては, 提案手法 4.3 の LDA による分類モデルの作成に対する実験 1, 4.4 のラベルとなる代表語を選出することに対する実験 2 と実験 3 に分けて行った. 実験 1, 実験 2 と実験 3 は独立したものとして, 訓練用に用いる文書データ, 及び評価基準は別々のものを利用した. 次節以降では, 実験に使用したデータの詳細と各評価実験の詳細についてを述べる.

5.2　実験に利用したデータ

実験で利用した文書データのコーパスについて説明を行う.

5.2.1　BCCWJ コーパス

本研究での実験用の文書データには, 国立国語研究所を中心として開発された, 現代日本語書き言葉均衡コーパス (BCCWJ コーパス) を利用する [14]. このコーパスは, 書籍, 雑誌, 新聞といった出版物をはじめ, ブログ, ネット掲示版のようなインターネット上の文章といった, 日本語の様々なレジスターにおける日本語の書き言葉をサンプルして, 文書構造や形態論情報を加えて TSV ファイルや XML ファイルの形式で収録したものである. このうちの XML ファイルに関しては, 各レジスター毎に発行年, ジャンル, 発行地域などの情報がサンプル ID を通じてデータに紐付けした状態で収録されている.

本研究では, コーパス中の XML ファイルからサンプルされた文章, 及び文章のジャンルやタイトルなど必要な情報を抽出して, 文書データとして紐付けてまとめるデータの前処理を行い, こうして得られた文書データをそれぞれの実験に用いた. 実験 1, 実験 2 と実験 3 に用いたデータは, レジスターや 1 件のデータにおける文章の長さが異なっている.

5.2.2　実験 1 に用いたデータ

実験 1 では, 日本十進分類法 (NDC) の第一次区分によって分類されている書籍レジスターのデータのうち, 2001 年から 2005 年までに出版された書籍からのサンプルで, 1 件につき 1000 文字前後の固定長で収録されているデータ 9575 件を使用した. 実験では, データの分類記号である数字をそのままデータのジャンル情報となる ID として利用している. 使用したデータの分類ごとの件数を表 5.1 に示す.

表 5.1: 実験 1 のデータ類目とデータ件数

分類記号	類目	データ件数
0	総記	329
1	哲学	545
2	歴史	859
3	社会科学	2,497
4	自然科学	1,030
5	技術	918
6	産業	437
7	芸術	653
8	言語	182
9	文学	2,125

5.2.3　実験 2, 実験 3 に用いたデータ

実験 2 及び実験 3 では, 実験 1 と同じ書籍レジスターのデータを用いるとトピックの話題が広義的であるため, 実験としての正解となる, トピックの代表語として選出されるべき語の設定が難しくなることを考慮して, 記事のタイトル部分を正解の候補として利用できる新聞レジスターのデータを使用した. 新聞レジスターのデータのうち, 2001 年から 2005 年までに出版された新聞からのサンプルで, 文章の長さは可変長でデータ 1 件に記事 1 つ分の文章が収録されているデータ 1117 件を利用した, 実験では, 新聞記事のタイトルの部分を別途抽出し, 実験 2 における正解とみなす単語群として利用した.

5.3　実験 1

実験 1 では, 最初にコーパス中の文書データに対して形態素解析を行って単語と重みの情報を取り出し, コーパス中のファイルから得られるジャンルの情報を, 単語を含む文章が持つ話題の情報として紐付けた. 単語が複数の文章に含まれていた場合は, 重みの一番大きい文章が持つ話題の情報を紐付けた. その後, 単語情報全体を 5 分割し, 分割されたうちの 4 つを分類モデルの訓練用, 残りの 1 つを評価実験でのテスト用として, 訓練用の部分のみで LDA に基づいて単語からその単語を含む文章が持つ話題を推定する分類モデルを作成し, テスト用の部分の単語に対して話題の分類を行って得られた結果が紐付けられた話題

の情報と一致するかどうかを正しく分類できているかどうかとして, 分類判定を行った. これをテスト用データに使用する部分を入れ替えながら 5 回実施する 5-fold cross-validation によって, 提案手法における LDA による分類モデルの分類の精度を測定した.

5.3.1 実験 1 の結果

実験 1 によって得られた, LDA による分類モデルの分類精度は 0.624 であった. この結果から, 提案手法は文章中の単語を 6 割程の精度で話題ごとに正しく分類することに成功していると考えられる. 分類モデルによって分類された ID 値を行, 実際に単語に紐付けられていた ID 値を列とした対応表を表 5.2 に示す.

加えて, 表 5.2 から分類モデルによって分類された ID 値が 9 に偏っていることが伺えるため, 詳細の調査を行った. まず, ID 値 9 に属している単語のデータに偏りがあるかを調べるため, ID 値 9 に該当するデータを取り除いた状態で再度実験を行ったが, その分類結果は ID 値 3 に偏るようになり, ID 値 3 も取り除いて実行すると ID 値 2 に偏るといったように, 分類結果が常にひとつの ID に偏るようになっていたため, 特定の ID 値に属するデータ自体に偏りが存在する様子は見られなかった. 続けて, ラベルとして用いている ID の数値に対して, 最大値や最小値に偏るようなアルゴリズム上の問題があるかを調べるため, データに割り振っていた ID 値を逆順にして再度実験を行ったが, 分類結果は ID 値が逆順になっただけで分類結果の分布に変化は見られなかったため, アルゴリズムがラベルの数値に依存している様子も見られなかった. 以上の結果から, 判別結果がひとつの ID 値に偏る原因は特定できなかったが, 実験上は問題ないことが確認できた.

表 5.2: 実験 1 の分類結果の対応表 (分類モデルによって分類された ID 値が行, 実際に単語に紐付けられていた ID 値が列)

	0	1	2	3	4	5	6	7	8	9
0	432	18	20	80	12	27	6	15	3	55
1	13	628	40	87	22	25	7	10	4	86
2	48	95	2039	177	44	53	37	38	9	247
3	64	122	181	2299	156	103	65	122	34	296
4	19	25	39	123	1037	35	15	11	5	63
5	21	16	48	118	49	952	15	20	4	70
6	13	14	31	78	29	30	381	11	3	40
7	18	43	52	88	45	35	14	957	2	95
8	11	23	34	70	26	10	4	13	263	31
9	174	283	340	490	512	471	244	359	103	2519

5.4　実験 2

実験 2 では, 新聞記事の文書データに対して形態素解析を行って単語と重みの情報を取り出し, 1 つの記事で使用されている全ての単語の集合を LDA による分類モデルの分類から得られたトピックと想定して用意した. また, 新聞記事のタイトルは記事の文章を要約したものであるという前提のもと, 単語群のうち記事のタイトル中に含まれる語をトピックの代表語として選出されるべき語として設定した. その後, 単語群から代表語として, TF-IDF による単語の各文章における重みの合計値が高い 3 つの単語, 及び 3 つの単語それぞれに対するコサイン類似度が高い 3 つの単語を付属させた最大 12 個の単語を重複を許して選出し, 選出された単語のうち 1 つでも代表語として選出されるべき語として設定した語が含まれていれば正解という基準で, 各単語群に対して正解か不正解かを評価していった.

5.4.1　実験 2 の結果

実験 2 での出力結果として, 以降に正解とみなされた例, 不正解とみなされた例 1 件ずつにおいて, データの概容と, 重みから選出された単語にコサイン類似度が高い 3 つの単語を付属させた, 4 つの単語を 3 通り示す. 正解とみなされた例には, ”学力検査 3 月 5 日来年度の公立高入試”と言う記事タイトルで, ”道教委は二十五日, 来春の入学者を選抜する二〇〇二年度道立高校入試日程について, 学力検査は〇二年三月五日, 合格発表は同十六日と発表した. 〜”といった入学試験の学力検査の日程を報じた記事がある. これに対し提案手法は, ”入試 学力 要項 選抜”, ”学力 要項 選抜 推薦”, ”入学 道立高校入試日程 願書受付”といった代表語を出力し, ”学力”, ”入試”といった記事タイトルにも含まれる単語を選出していたので, 正解とみなされた. 不正解とみなされた例には, ”三沢沖異常接近問題 海自機機長ら 3 人懲戒処分”と言う記事タイトルで, ”三沢市沖の太平洋で今年七月に海上自衛隊の哨戒機Ｐ３Ｃが降下し漁船に至近距離まで接近した問題で, 海上自衛隊は二十日, 同機が所属する第二航空群（八戸市）の当時の司令ら三人を注意などの懲戒処分にした. 〜”といった海上自衛隊員に対する懲戒処分を報じた記事がある. これに対し提案手法は, ”航空 同機 降下 古庄”, ”注意 距離 集団 太平洋”, ”司令 哨戒 至近 降下”といった代表語を出力したが, 出力した代表語の中に記事タイトルにも含まれる単語はなかったので, 不正解とみなされた. 実験 2 の全単語群に対する代表語選出の正解率は, 0.935 であった. この結果から, 提案手法は同じの話題に属する単語が集まった単語郡から, 9 割程の精度で代表語を適切に選出することに成功していると考えられる.

5.5　実験 3

実験 2 に関して, 新聞記事のタイトルに含まれる単語がその新聞記事から作成された単語群における代表語であるとみなしてよいかという, 実験 2 における正解基準に関する疑問が残ったので, 実験 2 の正解基準の妥当性を診断するための補助実験として実験 3 を行った.

実験 3 では, 実験 2 で使用した新聞記事の文書データから, 記事のタイトルと実験 2 によって選出された代表語及び実験 2 での正解判定の情報を取り除いた, 新聞記事の本文の

みの文書データをランダムに 30 件サンプルした．サンプルされた文書データを実験用のデータとして 1 件ずつ人の手で読み, 文中に存在する名詞の中からその記事の話題を象徴していると判断した単語を 3 つ選出し, これを代表語として選出されるべき語として設定した．その後, 実験 2 で選出された TF-IDF による単語の各文章における重みの合計値が高い 3 つの単語, 及び 3 つの単語それぞれに対するコサイン類似度が高い 3 つの単語を付属させた最大 12 個の単語を参照し, 選出された単語のうち 1 つでも代表語として選出されるべき語として設定した語が含まれていれば正解という基準で, 各単語群に対して正解か不正解かを評価していった．

5.5.1　実験 3 の結果

サンプルされた 30 件のうち, 実験 3 で正解と判断されたものは 26 件であった．この結果から, 人手で代表語となるべき正解の語を用意した場合でも 9 割程の精度で代表語を適切に選出することに成功しているので, 実験 2 における正解基準は妥当なものであったと考えられる．

しかし, データの中には, 実験 2 で正解とされていたが実験 3 では不正解とされたもの, 実験 2 で不正解とされていたが実験 3 では正解とされたものが存在したため, 詳細を調査した．調査の結果, 実験 2 で正解とされていたが実験 3 では不正解とされた例には, ”スポーツと健康痛みを知る体の異変知らせる危険信号”と言う記事タイトルで, ”小泉内閣の構造改革には「痛みを伴う」ことが強調されている．手術などの苦痛と不安に耐えれば必ず健康を回復するという見通しがあれば, 伴う痛みも我慢もできる．しかし, 〜”といった記事のタイトルと本文の大部分がスポーツの話に置き換えた例え話で, 東大教授が政治に対する批評を行っている記事がある．これに対し実験 2 では例え話の部分から, ”痛み 信号 スポーツ 異変”, ”スポーツ 異変 信号 楽しみ”, ”信号 異変 見通し この世”と選出されて正解とされていたが, 実験 3 では人手で”東大, 内閣, 構造改革”と選出され, 不正解とみなされた．また, 実験 2 で不正解とされていたが実験 3 では正解とされた例には, ”水霊 (８ ２) 第三章 月夜とウナギ (２ ３)”と言う記事タイトルで, ”少しずつ昭彦が身近になっていく．なによりも彼の, 気取りのなさが詩子には好ましかった．大学を卒業したら, いまアルバイトをしている会社で働くことにする．〜”といった記事のタイトルが連載されている小説の作品や章の名前で, 記事内容はその本文であるような記事がある．これに対し実験 2 において”昭彦 カカオ そうこう アイリッシュ・ウイスキー”, ”詩子 ひさこ 昭彦 真弓”, ”ボトル カカオ そうこう アイリッシュ・ウイスキー”と選出されて不正解とされていたものの, 実験 3 では人手で”昭彦, 詩子 楽器”と選出され, 登場人物名から正解とみなされた例が発見された．

このように, データの中には新聞記事のタイトルを実験での正解として利用するにはふさわしくない例もあったことがわかった．

5.6　考察

実験 1 の結果から, 多数の文章から 6 割程の精度で単語を背景に持つ話題ごとに正しい分類に成功していることがわかった．また, 実験 2 及び実験 3 の結果から, 背景に持つ話題

ごとに分類された単語群から代表語を 9 割程の精度で選出できることがわかった. これにより, 提案手法は複数のテキストから名詞を話題ごとに分類し, その中から代表となる語を選出することに対して有効であると考えられる.

その他に, 今回の実験で考慮しきれなかった問題として, 提案手法全体が統合されていない点がある. 本実験では提案手法のうち, LDA による分類モデルを作成して単語をトピックごとに分類する工程, 分類したトピックから代表語を選出する工程をそれぞれ独立したものとみなして, 異なるデータに対して異なる評価実験を行っていたため, 提案手法中の各工程それぞれの処理は有効に機能することが確認できても, それぞれの工程を統合した場合にもうまく動作するかは確認がとれていない. そのため, これまで提案手法内で独立して実装及び評価実験を行っていた各工程を併せて, 同じ文書データに適用できるようにして, 提案手法全体としての実装や評価を行っていくことが, 本研究の今後の課題となる.

第6章　まとめ

本研究では, 文章を構成する単語の中に知らない新語や専門用語が含まれているなど, いずれかの単語が示す意味を知らない, 理解できないことに起因して文章の読解が困難になる問題について取り上げた. そしてこの問題において, 単語の辞書的な定義を知ることはできなくても, 単語と同じ話題の文章で用いられる関連語, 単語の上位概念となる語など, 単語が属している話題の情報を大まかな意味や背景として知ることができれば, 文章全体が示す意味も大まかにつかむ事ができる点に着目した. 本研究では, この問題を解決するアプローチのひとつして, トピックモデルに基づいて文書データから文章中の単語をトピックとして話題ごとに分類し, 分類した単語群の中から単語の重みを参照して代表となる単語を選んで単語がどの話題に関するものであるのかを推定することで, 文章中の単語がどの話題に属しているかを示す手法を提案した. 提案手法は大きく分けて訓練用の文書データに対して形態素解析を行い, 文章を構成している単語の情報を得る工程, 取得した情報から LDA による分類モデルを作成し, 文章中の単語をトピックごとに分類する工程, 手法の推定結果として, 分類したトピックから, ラベルとなる代表語を選出する工程の 3 つで構成されている. このうちの LDA による分類モデルの作成工程と代表語の選出工程の 2 つに対して, それぞれ性能の評価実験を行った. 実験より得られた結果から, 提案手法は多数の文章から 6 割程の精度で単語を背景に持つ話題ごとの正しい分類に成功すること, 背景に持つ話題ごとに分類された単語群から代表語を 9 割程の精度で選出できることがわかった. 本研究の今後の課題として, これまで提案手法内で独立して実装及び評価実験を行っていた各工程を併せて, 同じ文書データに適用できるようにして, 提案手法全体としての実装や評価を行っていく必要がある.

謝辞

本研究を進めるにあたって, 丁寧にご指導を下さった新美礼彦准教授に深く感謝致します. また, 新美研究室の皆様, その他研究に関するアドバイスをしていただいた方々に深く感謝いたします.

参考文献

[1] 奥村 学, 高村 大也, 言語処理のための機械学習入門, コロナ社, 2010.

[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 参照 2016-1-25, http://taku910.github.io/mecab/

[3] rmecab, 参照 2016-1-25, https://sites.google.com/site/rmecab/

[4] Hofmann, T. (1999). "Probabilistic Latent Semantic Indexing". SI-GIR.

[5] Blei, D. M., Ng, A.Y. and Jordan, M.I. (2003). "Latent Dirichlet Allocation". Journal of Machine Learning Research, Volume 3, pp.993-1022.

[6] 岩田 具治, MLP 機械学習プロフェッショナルシリーズ トピックモデル, 講談社, 2015.

[7] The R Project for Statistical Computing, (online), 参照 2016-1-25, http://www.rproject.org/

[8] CRAN - Package MASS, 参照 2016-1-25, https://cran.r-project.org/web/packages/MASS/index.html

[9] CRAN - Package proxy, 参照 2016-1-25, https://cran.r-project.org/web/packages/proxy/index.html

[10] 新納 浩幸, 佐々木 稔 (2013). "k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応". 研究報告自然言語処理（NL）, 情報処理学会, pp.1-7.

[11] 白井 匡人, 三浦 孝夫 (2014). "トピックモデルに基づく文書ストリームのマルチラベル分類", DEIM Forum 2014 A9-1, pp1-5.

[12] 堀内 佑城, 輪島 幸治, 古川 利博 (2015). "ヘルプデスク作業効率化のためのラベリング自動化". DEIM Forum 2015 D1-4, pp1-4.

[13] 清水 浩平, 萩原 将文 (2006). "形容詞共起を用いた単語の印象推定法". 電子情報通信学会論文誌. D, 情報・システム, J89-D(11), 2483-2490.

[14] 概要 現代日本語書き言葉均衡コーパス (BCCWJ), 参照 2016-1-25, http://pj.ninjal.ac.jp/corpus_center/bccwj/

表 目 次