多義語に対するトピックモデルによる話題推定手法

田中　桂介†　　新美　礼彦†

† 公立はこだて未来大学大学院システム情報科学研究科　041–8655 北海道函館市亀田中野町 116 番地 2
E-mail: †{g2116024,niimi}@fun.ac.jp

あらまし　本研究は, 文章中の単語の属する話題の推定を補助する手法の提案を目的とする. 文章中の単語に対して, その単語が何に関する語であるのかを知る方法として, 単語の属する話題を推定することがある. しかし, そのような場合には, 同じ表記で複数の話題に用いられる多義語に対して, ひとつの話題にしか推定結果を振り分けることができない問題がある. そこで本研究では, 文章中の多義語を, その用例ごとに切り分けてグループとしてまとめる手法を提案した. 提案手法では, (1) ひとつの多義語に対して, その前後 2 つずつの名詞を単語のチャンクとして集計してまとめる, (2) 単語のチャンクに対してクラスタリングを行い, 用例ごとに切り分ける, という手順となっている. これによって, 文章中の多義語のひとつひとつを, 用いられる用例ごとに切り分ける. 提案手法によるクラスタリング結果に対する評価実験によって得られた結果から, 提案手法は多義語に対して 8 割程の精度で用例ごとの正しい切り分けに成功することがわかった.

キーワード　テキストマイニング, トピックモデル, LDA, 多義語, クラスタリング

1. はじめに

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, 特に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, 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.

そこで我々は, このような問題の解決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手法の提案を行った [1]. 提案した手法では, トピックモデルに基づいて文章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している. しかし, この手法には複数の用例を持っていて, 同じ表記で複数の話題に用いられるような多義語に対して, ひとつの話題にしか推定結果を振り分けることができず, 多義語の使い分けに対する正しい話題の分類が不可能である問題がある.

この問題点に対して, 多義語の用例ごとに異なる話題の推定結果の振り分けを可能にするため, 同じ表記を用いながら複数の話題で用いられる多義語を, その用例ごとに分類することを本研究の目的とする.

そこで本研究では, 多義語の前後 2 つずつの名詞を単語のチャンクとして集計してまとめ, 単語のチャンクに対してクラスタリングを行って多義語の用例ごとに切り分けることで, 文章中の多義語を用いられる用例ごとに切り分ける手法を提案する.

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す.

2. 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 [2]. 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.

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書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書の持つ話題を推定していく. また, この BoW 表現における文書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.

本研究における単語の属する話題の推定手法では, 文書にどのような単語が含まれているかの情報から文書の持つ話題を推定する工程を, 単語がどのような文書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させている.

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 [3]. 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, 本研究における手法はこのうち LDA のモデルを応用させる研究にあたる.

3. 関 連 研 究

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 [4]. 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 [5]. 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 [6]. 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 単語クラスタリングの語義判別問題への応用

佐々木らの研究に, 語義判別問題の解決を目的とした専用のシソーラス構築の自動化のため, 意味的につながりやすい単語を手がかりとしてクラスタリングを行う研究がある [7]. この研究では, 共起関係を持つ単語の共起データを抽出し, 潜在的意味解析を行ったうえで単語クラスタリングを実行することで, 意味的につながりを持つ単語同士によるクラスターの構築を実現させている.

3.5 関連研究との違い

本研究における単語の属する話題の推定手法の特色は, トピックモデルを単語単位に着目して適用させる点であり, 本研究で行う多義語の用例ごとの切り分けにおける特色は, N-gram を利用して単語のチャンクに対してクラスタリングを行う点である.

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, 本研究における手法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, 本研究における手法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, 本研究における手法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 とは文章中の単語において, その単語がどの語義を表しているかの判断を行う, 語義判別問題の解決を図る研究として目標が類似しているが, 本研究における手法では, 共起行列でなく N-gram を利用して単語のチャンクへのクラスタリングを行うため, 目標となる単語とより近くで共起する単語や, 共起する単語との前後関係を重視したクラスタリング結果が得られることが考えられる.

4. 提 案 手 法

4.1 提案手法概容

本研究の提案手法では, 文書データに対して, 以下の処理を行うことで文章中の多義語のひとつひとつを, 用いられる用例ごとに切り分けてグループとしてまとめる.

（1） ひとつの多義語に対して, その前後 2 つずつの名詞を単語のチャンクとして集計してまとめる

（2） 単語のチャンクに対してクラスタリングを行い, 用例ごとに切り分ける

本手法は, 多義語の用例ごとの切り分けにおいて, N-gram を利用して単語のチャンクに対してクラスタリングを行うため, 目標となる単語と近くで共起する単語や, その前後関係を重視したクラスタリング結果が得られることが特徴となる.

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 単語のチャンク集計

文書データに対して形態素解析を実行し, 分割された形態素のひとつひとつを単位とした 2-gram を作成して, そこから多義語が中心となる要素を取り出すことにより, データに出現する多義語ごとにその前後 2 つずつの単語をまとめた単語のチャンクを得る. ひとつの語に対して, それぞれ文章中でどの語と同時に出現しているかによって別の要素として集計していくことで, その語の様々な用例を得ることができる. 本手法では名詞

のみを 2-gram の作成における対象とする. 形態素解析の実行は, 形態素解析ソフト MeCab [8] 及び R 言語の RMeCab パッケージ [9] を用いて行う.

4.3 クラスタリングの実行

単語のチャンク集計によって得られたチャンクに対して, 多義語の持つ用例の数をクラスター数とした k-means 法による非階層のクラスタリングを行うことにより, チャンクを多義語の用例ごとに切り分ける. 文章中の多義語の出現例をまとめた単語のチャンクを用例ごとに切り分けることで, 同じ表記の単語に対してその用例に応じて異なる単語の話題の推定結果を与えることが可能となる. クラスタリングに用いるアルゴリズムは, Hartigan-Wong のアルゴリズム [10] を利用する. クラスタリングの実行は, R 言語の MASS パッケージ [11] を用いて行う.

5. 実　　験

5.1 実 験 概 容

提案手法の評価実験に関しては, 提案手法に基づいて実際に実験データに対する多義語の用例ごとの切り分けを行えるかを調べる実験 1, 実験 1 で得られた結果と人手による判断結果との比較によって妥当性を判断する実験 2 の 2 つを行った. 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 実験に用いたデータ

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 [12]. 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, データの XML ファイルから文章のみを抽出し, 新たにテキストファイルとして保存したうえで解析処理を行っている.

実験では, 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を, 分類区分となる総記, 哲学, 歴史, 社会科学, 自然科学, 技術, 産業, 芸術, 言語, 文学からそれぞれ 300 件, 合計 3000 件を使用した.

5.3 実　験　1

提案手法に基づいて実際にデータ中の多義語に対して正しく用例ごとの切り分けを行えるかを調べるため, データから多義語を含む単語のチャンク集計とクラスタリングの実行をする, 実験 1 を行った.

実験 1 では, 実験データの全体から 2-gram を作成し, そこから 41 の多義語に対して, 多義語が中心となる要素を取り出して, 多義語の前後 2 つずつの単語をまとめた単語のチャンクを得た. その後, 各多義語ごとが持つ用例の数をクラスター数とした非階層のクラスタリングを実行し, クラスタリング結果に対して手動でクラスターに用例を割り振った. 最後に実験の結果として, それぞれの単語のチャンクにおける用例を, そのチャンクを含むクラスターに対して割り振られた用例として集計した.

実験 1 における 41 の多義語は, SENSEVAL-2 日本語タスク [13] の辞書タスクにおける評価単語から選定を行った. また, クラスター数及びクラスターに対する振り分けに利用する多義語ごとの用例については, 日本語 WordNet [14] の WordNet 検索から得られた概念を参考に選定を行った.

5.4 実験 1 の結果

実験 1 によって, 41 の多義語それぞれに, 40 個から 4849 個までの単語のチャンクが取り出され, それらを 2 個から 4 個までの用例によって分類した結果が得られた. 結果の例として, ” 一方” と” 時代” の 2 つの多義語に対するクラスタリングの実行結果を表 1 に示す. ” 一方” のクラスター数は 2, ” 時代” のクラスター数は 3 である.

この後, 各多義語の単語のチャンクが実際に振り分けられたクラスターを見て, クラスターに属する単語のチャンクの傾向に適していると考えられる用例をクラスターに割り振っていった. 例として, ” 一方” では, ” 予防-強調-一方-様々-がん”, ” 者-保護-一方-瑕疵-軽微” といったチャンクが存在するクラスター 1 に” 標準または規格またはレベルまたはタイプまたは社会規範と一致する” の用例, ” 不可能-どちら-一方-謡-追従”, ” 反対-人-一方-こと-裁判” といったチャンクが存在するクラスター 2 に” 問題の 2 つの面の 1 つ” の用例を割り振った. 同様に, ” 時代” では, ” 経済-成長-時代-はじめ-頃”, ” 理性-啓蒙-時代-絶対-者” といったチャンクが存在するクラスター 1 に” 時間の量” の用例, ” プロ-現役-時代-スタイル-陸上”, ” ペース-現代-時代-ライブ-ラジオ” といったチャンクが存在するクラスター 2 に” 現代の環境と思想” の用例, ” 傾向-明治-時代-遼東-還付”, ” 環境-徳川-時代-鎖国-政策” といったチャンクが存在するクラスター 3 に” 顕著な特徴を持っている歴史の時代” の用例を割り振った.

表 1 からは, どちらの多義語も単語のチャンクに対する分類の結果がどれか 1 つのクラスターに偏っている傾向が見てとれる. この傾向に関しては, 程度に差はあれど全ての多義語において共通していた傾向であり, これがデータ内に似た用例が多いからというデータが原因となって起こっているものであるのか, それともクラスタリングによる分類がうまく動作していないことが原因となって起こっているものであるかを調査する必要が生じた.

表 1　実験 1 によるクラスタリングによる分類結果例

多義語	クラスター 1	クラスター 2	クラスター 3	合計
一方	24	99	-	123
時代	48	774	2	824

5.5 実　験　2

実験 1 で得られたクラスタリングの分類結果が妥当なものであるかの判断を行うため, 正解となる単語のチャンクに対する用例を手動で用意して, クラスタリングの分類結果と比較する,

実験 2 を行った.

実験 2 では, 実験 1 で得られた, 41 の多義語の前後 2 つずつの単語をまとめた単語のチャンクのうち, 各単語 40 個ずつをクラスタリングによる分類結果の情報を取り除いた状態で抽出した. その後, 抽出された単語のチャンクに対して, 1 件ずつ人の手で読み, 多義語の用例から最もふさわしいと判断できたものに振り分けていった. その後, 人手で判断されて振り分けられた用例と, 実験 1 によって得られた単語のチャンクにおける用例を比較し, 一致していれば実験 1 での切り分けは正解であるという基準で, 各単語のチャンクに対して正解か不正解かを評価し, それぞれの多義語に対して正解率を集計した.

実験 2 での人手での振り分けに利用する多義語の用例は, 全て実験 1 で利用したものと同じである.

5.6 実験 2 の結果

実験 2 によって得られた, 41 の多義語の正解率の平均は 0.76 であった. 結果の例として, 実験 1 の結果と同様となる" 一方" と" 時代" の 2 つの実験 1 によるクラスタリングでの用例の振り分け結果と実験 2 による人手での用例の振り分け結果の対応表を表 2, 及び表 3 に示す. " 一方" における正解率は最高値となる 0.85, " 時代" における正解率は最低値となる 0.53 であった.

表 2 からは, 人手による振り分けられた結果が実験 1 における単語のチャンクに対する分類の結果と同様に 1 つのクラスターに偏っている傾向が見られた. 同様に正解率の高い多義語は, どれもクラスタリングによる分類結果に合わせて人手での振り分け結果が偏っていたため, 実験 1 で見られた単語のチャンクに対する分類の結果がどれか 1 つのクラスターに偏っている傾向は, データが原因となって起こっているものであったと考えられる.

表 3 からは, クラスタリングでの用例の振り分け結果が不正解であると判断された例の中では, 特に" 安土-桃山-時代-竹島-家" や, " 所蔵-室町-時代-末期-伝" のような, 人手によってクラスター 3 に割り振られた" 顕著な特徴を持っている歴史の時代" の用例で用いられている単語のチャンクに対して, 実験 1 で誤ってクラスター 1 に割り振られた" 時間の量" の用例が振り分けられた例がほとんどであることが見てとれる. 同様に, 正解率の低い多義語では, どれも実験 1 で分類された件数が極端に少ないクラスターに振り分けられた用例のチャンクが, 人手で判断した場合は件数が多かった, という傾向が見られた.

表 2 実験 2 による" 一方" に対する分類結果の対応表 (クラスタリングによって分類された用例が行, 人手によって振り分けられた用例が列)

	1	2
1	4	1
2	4	31

5.7 考　　察

実験 2 の結果は, 文章中に出現する多義語に対して, その前後に出現する語をチャンクとしてまとめることで, その多義語がどの用例で用いられているのかを 8 割程の精度で判断することが可能であることを示す. したがって, 提案手法は文章中の多義語に対して 8 割程の精度で用例ごとの正しい切り分けに成功することがわかった. これにより, 多義語をその用例ごとに切り分けて, 別の単語として扱うように前処理を施してから単語の属する話題を推定することで, 多義語に対する話題の推定性能が向上することが期待できる.

表 3 実験 2 による" 時代" に対する分類結果の対応表 (クラスタリングによって分類された用例が行, 人手によって振り分けられた用例が列)

	1	2	3
1	3	0	15
2	2	17	2
3	0	0	1

その他に, 今回の実験で考慮しきれなかった問題として, 事前にクラスタリングによる分類を行うべき多義語やその用例を用意しておく必用がある点がある. 現状では, 単語のチャンク集計及びクラスタリングの実行を行う対象となる多義語については全て事前に人手での調査や設定を行う必要があり, それに応じてクラスタリングにおけるクラスター数となる多義語の用例の数も事前に決めておく必用がある. 加えて, 分類後のクラスターに対してどのクラスターがどの用例に該当するものであるかも手動で割り振っている. このように, 現段階での提案手法は手動での作業に頼っている部分が大きく, 膨大な種類の多義語を含む実データに対する実行は厳しいものとなっている. そのため, 手法の対象となる多義語やその用例を辞書から取得するといった, 手法の実行に必要な動作を自動化させていくことが, 本研究の今後の課題となる.

6. 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の解決を補助する方法のひとつとして提案した, 文章中の単語がどの話題に属しているかを示す手法が持つ, 多義語への動作がうまくいかない問題を取り上げた. そしてこの問題に対して, 多義語の使い分けに対応した話題の推定結果の振り分けを可能にするため, 文章中の多義語を用いられる用例ごとに切り分ける手法を提案した. 提案手法は文書データに出現するひとつひとつの多義語に対して, その前後 2 つずつの名詞を単語のチャンクとして集計してまとめる工程, 集計された単語のチャンクに対してクラスタリングを行い, 用例を切り分ける工程の 2 つで構成されている. また, 提案手法に対して, 人手で単語のチャンクに対して割り振った用例を正解として, 提案手法に従って切り分けられた用例との比較を行う評価実験を行った. 実験より得られた結果から, 提案手法は文章中の多義語に対して 8 割程の精度で用例ごとの正しい切り分けに成功することがわかった. 本研究の今後の課題として, 現状では手動での作業に頼っている, 手法を実行する対象となる多義語やその用例の取得を自動化させていく必要がある.

文　　献

[1] 田中 桂介, 新美 礼彦 (2016). "トピックモデルによる単語の属する話題分類と代表語抽出", DEIM Forum 2016 C2-6, pp.1-6.

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

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

[4] 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.

[5] 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, pp.1-5.

[6] 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, pp.1-4.

[7] 佐々木 稔, 新納 浩幸 (2003). "単語クラスタリングの語義判別問題への応用". 情報処理学会研究報告自然言語処理（NL）. 154-21, pp.145-142.

[8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 参照 2017-1-14, http://taku910.github.io/mecab/

[9] rmecab, 参照 2017-1-14, https://sites.google.com/site/rmecab/

[10] J. A. Hartigan and M. A. Wong (1979). "A K-Means Clustering Algorithm". Journal of the Royal Statistical Society. Series C (Applied Statistics). Vol. 28, No. 1, pp. 100-108.

[11] CRAN - Package MASS, 参照 2017-1-14, https://cran.r-project.org/web/packages/MASS/index.html

[12] 概要 現代日本語書き言葉均衡コーパス (BCCWJ), 参照 2017-1-14, http://pj.ninjal.ac.jp/corpus_center/bccwj/

[13] 黒橋 禎夫, 白井 清昭 (2001). "'SENSEVAL-2 日本語タスク", 電子情報通信学会言語理解とコミュニケーション研究会, NLC36-48, pp.1-8.

[14] 日本語 WordNet, 参照 2017-1-14, http://compling.hss.ntu.edu.sg/wnja/