可読性に基づいた日本語テキスト情報の特徴量評価

Evaluation of Japanese Text Information Features Based on the Readability

輪島　幸治† 木暮　啓†† 古川　利博††† 佐藤　哲司††

† 筑波大学 図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
††† 東京理科大学 工学部情報工学科　〒 125-5846 東京都葛飾区新宿 6-3-1
E-mail: †kwajima@ce.slis.tsukuba.ac.jp, ††{kei.kogure@dentsu.co.jp,satoh@ce.slis.tsukuba.ac.jp}

あらまし　近年，情報化社会の発展に伴い，WEB 上のテキスト情報の蓄積，流通が増加してきている．テキスト情報の増加に伴い，利用者に有用な情報推薦が望まれている．現状の情報推薦では，嗜好，内容ベース，個人属性，利用状況が用いられているが，推薦されるテキスト情報の内容の読みやすさは十分に考慮されていない．そこで，読みやすさに基づいた情報推薦を目的に，特徴量の評価手法を提案する．評価対象の特徴量には，表層情報，話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現の 10 種類，31 個，2,071 次元の特徴量を用いる．提案手法では，特徴量行列を NMF で行列分解し，重要な基底の重み付け係数評価および特徴量の寄与率を評価する．提案手法を実装評価し，読みやすさに基づいた特徴量の評価が行えることを明らかにした．評価結果を報告する．

キーワード　NMF，LSI，LDA，評価表現，拡張固有表現

1. はじめに

近年，情報化社会の発展により，WEB 上のテキスト情報は企業の評判分析 (注1) を始め，多様な分野で活用されている．しかし WEB 上のテキスト情報は日々増加し，情報過多の状況が発生している [1]．よって，利用者により良い情報推薦が望まれている．さて昨今，SIPS（Sympathize Identify Participate Share & Spread）[2] など，消費者間コミュニケーションは購買行動に大きな影響を与えることが知られている (注2)．SIPS の消費者間コミュニケーションはメディアで行われる．ここで，メディアとは，コミュニケーションを秩序づけるメカニズムの一つである．伝播メディアと成果メディアに大別される．

伝播メディアとは，コミュニケーションが行われるメディアや状況全般である．ソーシャルメディア (注3) (注4) (注5) や，オンラインコミュニティ (注6) (注7) (注8)，チームコミュニケーションツール (注9)，エンタープライズ向けのソーシャル・ネットワーク・サービス (注10) などがある．成果メディアとは，文脈や意味を接続するメディアである．自己と他者，行為と体験で定義されている [3]．代表的なコミュニケーション・モデルには，D.K. バーロの SMCR モデルがある [4]．SMCR モデルでは，コミュニケーションの要素は，「送り手 (Source)」「メッセージ (Message)」「チャネル (Channel)」「受け手 (Receiver)」で定義される．メッセージは符号化された記号である．記号の伝搬・解釈は，伝播メディアであるチャネルで行われる．伝播メディアは受け手の基本属性（世代・社会的関心・消費行動など）で異なる (注11) (注12) (注13)．ここで，C.S. パースの記号論では，記号表現は，「記号 (sign)」「解釈項 (interpretant)」「指示対象 (object)」の三項関係で定義されている [5]．記号は，イコン (icon:類像)，インデックス (index:指標)，シンボル (symbol:象徴) に大別される [6]．記号表現が文書の場合，記号は文章 (シンボル)，指示対象は記号を代替する対象 (内容) である．解釈項は記号を理解する要素であり，読みやすさなどがある．

ところで，コミュニケーションで受け手に情報を伝達するには，記号表現で指示対象を伝えるメッセージが必要である．文書を始めとしたテキスト情報のコミュニケーションでは，書き手 (送り手) が，伝播メディア (チャネル) の記号表現 (メッセージ) で，読み手 (受け手) に伝達する必要がある．記号表現で指示対象が読み手に伝達できない場合，記号表現が欠損している，記号が適切でない，解釈項が十分でない場合などがある．また，コミュニケーションの伝播メディアが異なる場合などもある．

(注1)：Oracle Social Relationship Management Cloud：
http://docs.oracle.com/cloud/latest/srmcs_gs/index.html
(注2)：Klout： https://klout.com/
(注3)：Twitter： https://twitter.com
(注4)：Instagram： https://www.instagram.com
(注5)：Youtube： https://www.youtube.com/
(注6)：Apple サポートコミュニティ：
https://discussionsjapan.apple.com
(注7)：Cisco - Support Community：
https://supportforums.cisco.com/ja
(注8)：Intel - Support Communities：
https://communities.intel.com/community/tech
(注9)：Slack： https://slack.com
(注10)：Yammer： https://www.yammer.com

(注11)：2015 年 国民生活時間調査：
https://www.nhk.or.jp/bunken/research/yoron/20160217_1.html
(注12)：週間高世帯視聴率番組 10 タイムシフト視聴率 - ビデオリサーチ：
https://www.videor.co.jp/tvrating/timeshift/index.html
(注13)：JNN データバンク（生活者実態調査）：
http://jds.ne.jp/datebase01.html

昨今の伝播メディアでは，情報推薦が行われている．既存の情報推薦では，嗜好，内容ベース，個人属性，利用状況が用いられている [7]．しかし，既存の内容ベースの情報推薦では，話題や評価表現など記号の特徴量は用いられているが，解釈項である読みやすさなどは十分に考慮されていない．これは，特徴量の多くが読みやすさなど解釈項に基づく評価が十分にされていないことに起因している．解釈項が十分でない場合，コミュニケーションの受け手には，指示対象 (内容) が伝達されない．したがって，伝播メディアの種別や基本属性に関わらず読みやすさは重要なコミュニケーション要素である．そこで本研究では，読みやすさに基づいてテキスト情報の特徴量の評価を行う．

提案手法では，第一に，既存研究の手法および辞書を用いて，評価対象のテキスト情報の特徴量を抽出する．評価する特徴量は，表層情報，話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現の 10 種類，31 個，2,071 次元の特徴量である．第二に，既存手法を用いて特徴量変換を行う．最後に，提案手法の基底選択および特徴量評価を用いて，読みやすさに重要な特徴量を明らかにする．以下，本論文では 2 章で本研究の対象に関して述べる．3 章で関連研究を述べ，4 章で提案手法である特徴量の評価手法を記す．5 章では，実験結果と考察を行い，最後に 6 章でまとめと今後の課題を示す．

2. 用語の定義

2.1 環境白書

本研究では，読みやすさの評価対象に，環境省発行の年次報告書の環境白書を用いる (注14)．環境白書は，環境状況の報告と環境保全に関する施策で構成されている．最新の環境白書は，循環型社会白書や生物多様性白書と合本し，環境・循環型社会・生物多様性白書となっている．環境白書は，SDGs（Sustainable Development Goals）など，最新の国際的な潮流も取り上げている (2017 年)．また，環境白書の普及啓発冊子として，英語版の環境白書や図で見る環境白書，こども環境白書も発行されている．環境問題は社会的関心が高く，高齢化に次ぐ社会問題として 47%が関心を持っている．(注15)．昨今では，企業価値の評価プロセスに，環境（Environment），社会（Social），ガバナンス（Governance）という ESG 要素が組み入れられている．金融機関の融資では，融資先企業の環境経営の取り組みや環境配慮活動を評価し，融資の実行を判断する環境格付融資が行われている (注16)．したがって，企業は環境問題への取り組みとして，展示会 (注17) やグリーン電力証書 (注18) などで活動を展開している．

(注14)：環境省 環境白書・循環型社会白書・生物多様性白書：
http://www.env.go.jp/policy/hakusyo/
(注15)：電通総研 - 環境コンシューマー調査 2014：
http://dentsu-ho.com/articles/1011/
(注16)：環境省 総合環境政策：
http://www.env.go.jp/policy/kinyu/kakuzukeyusi_sokusin.html
(注17)：エコプロ 2017 環境とエネルギーの未来展：
http://eco-pro.com/2017/
(注18)：グリーン電力証書活用ガイド：
http://www.env.go.jp/earth/ondanka/greenenergy/

一方で，グリーンウォッシュと呼ばれるうわべだけやごく一部のみ環境配慮した行動には，厳しい視線が向けられているのも事実である [8]．よって，適切な環境配慮行動や，施策の理解のために環境白書は有用である．しかし，環境白書では,「カーボンバジェット (2017 年 P.33)」など専門的な語彙が用いられている．そのため専門分野の実務者・研究者・学生以外の読み手には理解が困難である．対して，普及啓発冊子のこども環境白書は，当該年度の環境白書の代表的な環境問題をわかりやすく解説している．こども環境白書では,「地球がどんどんあたたまる (2016 年 P.3)」など理解しやすい表現が用いられている．

そこで本研究では，こども環境白書が同じ年度の環境白書の代表的な環境問題を取り上げていることに着目し，環境白書とこども環境白書をテキスト情報として用いる．本研究では，読みやすい環境白書をこども環境白書，読み辛い環境白書を環境・循環型社会・生物多様性白書とし読みやすさの評価対象とする．

2.2 特徴量

本研究では読みやすさを評価するため，テキスト情報より，特徴量を抽出する．テキスト情報は言語に応じて形態的特徴などが異なる [9]．本研究では，評価に日本語の環境白書のテキスト情報を用いる．既存研究で，重要度評価に用いられているテキスト情報の要素は，下記の 6 つである [10]．

（1） 単語の重要度の値（TF-IDF 法など）
（2） 文の位置情報
（3） テキスト中のタイトル情報
（4） テキスト中の手がかり表現
（5） 文あるいは単語間のつながり情報
（6） 文間の関係を解析したテキスト構造

本研究では，一部の特徴量をアルゴリズムで抽出する．使用するアルゴリズムで特徴量の抽出行う際，本研究では，テキスト情報を Bag-of-words として取り扱う．Bag-of-words では，単語の出現頻度に着目し，文法や文脈，単語順は考慮しない．よって，(2)(3)(5)(6) は対象外とした．また,（1）単語の重要度の値は，複数の話題からなるテキスト情報では問題を生じることが指摘されている [10] [11]．そこで，本研究のテキスト情報の特徴量は（4）テキスト中の手がかり表現を用いる．手がかり表現とは，テキスト情報に含まれる重要な文を明示する語である．特徴量抽出の研究は過去数年にわたり，いくつか研究がなされてきた．したがって，畳語 [12] など個別の特徴量抽出の研究は数多く行われている．

一方で，ニクラス・ルーマンの社会システム理論に，ゼマンティクという概念がある [13]．ゼマンティクとは，高度に一般化され状況に依存せずに使用できる意味である．そこで，手がかり表現の特徴量抽出には，従来研究で作成された既存の辞書を用いる．よって，評価に用いる特徴量は，すべて既存研究の特徴量である．著者らが知る限り，日本語テキスト情報に関する既存の特徴量の辞書は 21 個ある．本研究では，既存の辞書に表層情報およびアルゴリズムの特徴量を加えた 31 個の特徴量を用いる．特徴量は，メルヴィル・デューイの十進分類法 [14] に基づき，表層情報，話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現に分類した．

ここで，特徴量は，C.S. パース記号論の記号表現に包含される．よって，解釈項を介し，受け手の感覚器官で収集され，処理される [15]．対して，感覚器官における感覚の大きさは，刺激の強度の対数に比例する [16]．したがって，テキスト情報に関わらず，記号の種類や数，次元数が多いことは，解釈項の評価には直結しない．また，情報の伝達効率や計算量の課題，解釈の恣意性や流動性もある．しかし，その一方で，ニクラス・ルーマンの社会システム理論に，複雑性の縮減という概念がある [17]．複雑性の縮減とは，複雑な環境に対応するため，対応する側が環境を上回る複雑性を内包することである．本研究では，解釈の複雑性を記号の種類や数，次元数で内包する．評価対象の記号表現は日本語のテキスト情報，記号は文章 (シンボル)，評価する特徴量は 10 種類，31 個，2,071 次元である．

3. 関 連 研 究

3.1 可読性評価に関する研究

テキスト情報の読みやすさに関する研究は，数多くなされている [18]．読みやすさの研究は語彙，文章の構造，クローズ法，リーダビリティの研究などがある．本研究は，リーダビリティの研究に包含される．リーダビリティの研究は，内容の理解のしやすさ（可読性），活字の読みやすさ（視認性），興味の度合いによる読みやすさに大別される．本研究は，可読性の研究である．リーダビリティの既存研究では，語数やひらがなの割合，単語数や音節数を用いている [19] [20] (注19)．テキストの自動生成 [21] や学術本の難易度推定 [22] の応用研究がある．本研究では，読みやすさに基づいた情報推薦が目的である．既存のリーダビリティの特徴量を用いた情報推薦で，読みやすさに基づいた情報推薦は行える．しかし，リーダビリティ研究の特徴量では，単語の頻度や比率，文書の長さなどの表層情報で評価されている．よって，話題や手がかり表現 (意味) は十分に考慮されていない．そこで本研究では，読みやすさの評価に話題，意味を包含する．特徴量を評価することで，新たな特徴量をリーダビリティの特徴量に採用できる．議論に先立ち，本研究で用いる既存の特徴量抽出手法の簡単な説明を行う．

3.2 テキスト情報の特徴量に関する研究

既存研究の特徴量抽出について簡単な説明を行う．テキスト情報から各種特徴量を抽出する方法には，表層情報，アルゴリズム，既存研究の辞書を用いる抽出手法がある．表層情報は，文の数，読点・句点の数，文の長さ，文字種類 [23]，読点間の距離 [24]，漢字包含率 [24] などがある．媒体分析 [25] や書き手評価 [26]，文献の判定 [27] の応用研究がある．アルゴリズムは，似た語彙の集合である話題を抽出する研究がある [28]．話題抽出のアルゴリズムには，LSI（Latent Semantic Indexing）[29] や LDA（Latent Dirichlet Allocation）[30] がある．LSI は文書分類 [31]，LDA はユーザ推薦 [32] の応用研究がある．

意味の特徴量を抽出する既存研究の辞書の研究を説明する．意味の特徴量は 8 種類に大別される．語種の特徴量は和語や漢語，外来語など語彙を分類した種類である．対象読者層の年代差評価 [33] の応用研究がある．基本語の特徴量は使用率が大きく，使用範囲が広い語彙である [34]．ニュースの語彙分析の応用研究がある [35]．意味属性の特徴量には，意味分類コードがある．意味分類コードは語を意味に基づいて分類したコードである [36]．文書の自動分類 [37] の応用研究がある．言語表現の特徴量は，機能語や複合辞などの機能表現がある [38]．価値判断の解析 [39] の応用研究がある．文末表現は，日本語の文章の固定化された文末表現 [40] [41] である．ベストアンサーの推定 [42] や，文書の内容分析 [43] の応用研究がある．品詞は，日本語のテキスト情報を形態素解析した結果の情報である．名詞比率や MVR，品詞構成率がある [44]．固有表現は，テキスト情報に含まれる固有名詞である．品詞体系に基づく手法と固有表現分類に基づく手法がある．品詞体系に基づく手法には，日本語形態素解析器 MeCab [45](注20) で用いられている IPADIC 辞書がある．固有表現分類に基づく手法には，MUC(注21) および IREX(注22) の規定に基づき拡張された拡張固有表現階層 [46] がある．固有表現クラス分類 [47] の応用研究がある．評価表現は，テキスト情報の評価を表す表現である [48]．既存研究には，日本語評価極性辞書 [49] [50] や単語感情極性対応表 [51]，評価値表現辞書 [52] がある．日本語評価極性辞書は，リスクの見積もり [53] や偏向性を可視化 [54] の応用研究がある．単語感情極性対応表は，感情推定 [55] の応用研究がある．評価値表現辞書は，トラブルを表す文の抽出 [56] の応用研究がある．

3.3 特徴選択に関する研究

複数の特徴量から，応用課題に有用な特徴量を選択する問題は特徴選択問題に分類される (注23)．本研究の応用課題は読みやすさである．特徴選択の既存手法は，ベースラインアプローチやモデルベースによる特徴量選択手法，再帰的特徴削減手法などがある．ベースラインアプローチは閾値を満たさないすべての特徴量を削除する手法である．不要な特徴量を削除し，処理を効率化するために用いられている [57]．モデルベースによる特徴量選択は，既存研究のモデルを用いて特徴量の重要性を評価する方法である．交換モンテカルロ法とマルチヒストグラム法を用いる方法 [58]．また，決定木 [59] や SVM [60] を用いる方法がある．再帰的特徴削減手法はモデルベース特徴量選択とベースラインアプローチを繰り返す手法である．本研究はモデルベースによる特徴量選択に関連した研究である．

4. 提 案 手 法

4.1 概　　要

本研究における提案手法の説明を行う．本研究はモデルベースによる特徴量選択に関連した研究である．特徴量選択に用いるアルゴリズムは，非負値行列因子分解（NMF:Nonnegative Matrix Factorization）[61] を用いる．NMF は，観測行列 Y を基底行列 H と係数行列 U の積に行列分解する手法である．

(注19)：Flesch-Kincaid readability tests：
https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

(注20)：MeCab：http://taku910.github.io/mecab/

(注21)：MUC-6：https://cs.nyu.edu/cs/faculty/grishman/muc6.html

(注22)：IREX：http://nlp.cs.nyu.edu/irex/

(注23)：特徴選択：https://ja.wikipedia.org/wiki/特徴選択

提案手法は，NMF による行列分解の結果を用いた，重要な基底の選択方法，重要な特徴量の評価方法で構成されている．本研究は，既存アルゴリズムの NMF をテキスト情報の特徴量の重要度評価手法に応用するという特徴を有している．また，既存のモデルベースの特徴量選択手法とは異なるアルゴリズムを用いている．加えて，既存研究に用いられている表層情報，話題だけでなくテキスト情報の意味を，既存研究の辞書を組み合わせて包含する．さらに，テキスト情報の記号表現の複雑性を内包するため，本研究では既存研究とは異なる特徴量も包含して，評価を行う．評価に用いる特徴量は，表層情報，話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現の 10 種類，31 個，2,071 次元の特徴量である．

本章では，まず，評価に用いる既存研究の特徴量を説明する．その後，特徴量変換手法である NMF について説明する．そして，提案手法の重要な基底の選択方法を説明し，最後に，重要な特徴量の寄与率を評価する方法を説明する．

4.2 準　　備

本研究で評価に用いる既存研究の特徴量を説明する．既存研究の特徴量抽出方法は，抽出方法ごとに表層情報，アルゴリズム，既存研究の辞書の 3 グループを定義した．各種抽出方法を説明する．

表層情報は 8 種類の特徴量より構成される．特徴量の次元数は 22 である．表 1 に本研究で用いる表層情報を示す．なお，本研究の文字種は，ひらがな，カタカナ，漢字，アルファベット，数字，半角記号，空白記号，全角記号である．

表 1　表 層 情 報

特徴量名	次元数	値の定義/例
文の数	1	整数値 (1 文中の [。][?][!] の合計頻度)
読点	1	整数値 (1 文中の [、] の頻度)
句点	1	整数値 (1 文中の [。])
文の長さ	1	整数値 (1 文中の総バイト数)
文字種 (頻度)	8	整数値 (ひらがな，カタカナ，漢字など)
文字種 (比率)	8	実数値 (同上)
読点間距離	1	実数値 (算出 文献 [24])
漢字含有率	1	実数値 (算出 文献 [24])

本研究では，アルゴリズムで話題の特徴量を抽出する．話題の特徴量は，潜在意味解析とトピックモデルの 2 種類より構成される [28]．抽出アルゴリズムは，潜在意味解析に Latent Semantic Indexing(LSI) [29] トピックモデルに Latent Dirichlet Allocation(LDA) [30] を採用する．各アルゴリズムに関して，簡単な説明を行う．

LSI は，行列 N を低ランク行列の積 $\mathrm{U^T H}$ にフロベニウスノルムが最小になるように行列分解する手法である．フロベニウスノルムは要素を二乗して総和をとった値の平方根である．行列 N を文書番号 D と語彙 V とした際の，LSI を式 (1) で示す．

$$\|N - U^{\mathrm{T}} H\|^2 FRO = \sum_{d=1}^{D} \sum_{v=1}^{V} \left(N_{dv} - u_d^{\mathrm{T}} h_v\right)^2 \qquad (1)$$

$U = (u_1, \cdots, u_D)$ は重み付き係数行列であり，K 行 D 列の実数行列である．$H = (h_1, \cdots, h_V)$ は語彙行列であり，K 行 V 列の実数行列である．ここで，K は低ランク行列の次元数である．LSI の行列分解には，特異値分解が用いられている．本研究の LSI の実装は，分かち書き (注24) されたテキスト情報に，Gensim [62](注25) を適用する．よって，特異値分解は，乱択特異値分解 [63](注26) を用いる．

LDA は，文書 w_d が低ランク行列 ϕ と θ をパラメータとして持つカテゴリ分布から生成されると仮定する手法である．全体集合を W，文書数を $d = (1, \cdots, D)$，トピックを $k = (1, \cdots, K)$ とした際の，文書 w_d の生成確率を式 (2) で示す．

$$p(w_d|\theta_d, \Phi) = \prod_{n=1}^{N_d} \Sigma_{k=1}^{K} p(k|\theta_d) p(w_{dn}|\phi_k) \qquad (2)$$

パラメータ θ_d はトピック分布，パラメータ Φ は単語分布集合である．また，N_d は文書 d に含まれる単語数，w_{dn} は文書 d の n 番目の単語，ϕ_k はトピック k の単語分布を示す．本研究における LDA の実装に関しては，LSI と同様に Gensim [62] を用いた．よって，各文書のトピック分布 θ_d の推定は，変分ベイズ法 [64] を用いる．

本研究の LSI の次元数や LDA のトピック数である K は，任意で値の設定を行う．LSI の K の最適なパラメータは 300 から 500 の範囲が提案されている [65]．本研究では K=300 を設定した．LDA に関しても同様に K=300 を設定した．よって，特徴量の次元数は 600 である．話題の特徴量を表 2 に示す．

表 2　話　　題

特徴量名	次元数	値の定義
潜在意味解析 (LSI) [29]	300	文書の重み付き係数 u_D
トピックモデル (LDA) [30]	300	文書のトピック分布 θ_d

LSI では，低ランク行列分解された重み付き係数行列 U の値によって，文書 D における話題 K の話題の重み付きが明らかになる．よって，LSI の特徴量には，各文書の K の重み付き係数を用いている．LDA では，各文書における θ_d を推定することで，話題の割合が明らかになる．よって，LDA の特徴量には，各文書の θ_d を用いている．表 2 の特徴量の値は実数である．

既存研究の辞書では，意味情報の特徴量を抽出する．意味情報の特徴量は，21 種類の特徴量より構成される．表 3 に既存研究の辞書による特徴量を示す．既存研究の辞書は，付録に一覧表を記載した．「定義/付与タグの例」列に記載がない特徴量の値は整数である．特徴量の次元数は 1,449 である．各特徴量を水平方向に連結し，評価を行う特徴量行列の作成を行う．特徴量行列の作成は，Pandas(注27) および Numpy(注28) を用いた．

(注24)：一定の規則にしたがって，区切られた文章を分かち書きと呼ぶ．
(注25)：gensim： https://radimrehurek.com/gensim/
(注26)：Blei Lab： https://github.com/blei-lab
(注27)：Pandas： http://pandas.pydata.org/
(注28)：NumPy： http://www.numpy.org

表 3　意 味 情 報

特徴量名	次元数	語数	値の定義/付与タグの例
語種	7	6519	和語，漢語，外来語
基本語 (1)	2	697	基本語二千に選定
基本語 (2)	6	6519	外国語学習基本語彙
基本語 (3)	2	424	基本語二千・六千に選定
意味分類コード (1)	233	697	体の類抽象的関係
意味分類コード (2)	487	6519	(同上)
意味分類コード (3)	307	424	(同上)
機能表現	122	29262	O，B-判断
文末モダリティ	32	32	可能性，表出
質問文末表現	38	38	質問，回答
IPA 品詞	14	-	連体詞，接頭詞，名詞
名詞比率	1	-	実数 (算出 文献 [44])
MVR	1	-	実数 (算出 文献 [44])
固有名詞	4	-	実数 (算出 文献 [45])
拡張固有表現	132	18075	数値表現
評価極性情報			
（用言編）	4	5280	ネガ（経験）
（名詞編）	51	13314	～になる（状態）客観
感情極性		110250	
（頻度）	2	-	ポジティブ・ネガティブ
（比率）	2	-	実数 (算出 文献 [48])
（平均値）	1	-	実数 (算出 文献 [48])
評価値表現	1	5234	-

4.3　非負値行列因子分解 (NMF)

特徴量変換のアルゴリズムについて簡単な説明を行う．アルゴリズムに用いる NMF は既存研究の手法である [61]．NMF は観測行列 Y を基底行列 H と係数行列 U の積に分解するアルゴリズムである．ここで，NMF の観測行列 Y の要素は，全て非負値 (≧ 0) であることが前提である．本研究では，観測行列 Y を 0 から 1 の間に正規化し，各変数の計測尺度の違いを考慮するため，L1 正則化による制約補正を行った．よって，分解結果の基底行列 H および係数行列 U も非負値となる．

NMF では，観測行列 Y が，相関行列や分散共分散行列であることを前提としていない．よって，主成分分析 (Principal Component Analysis) や因子分析 (Factor Analysis) などと異なり，観測行列 Y の各次元の特性に依存せず適用できる．本研究の観測行列 Y は，節 4.2 の各特徴量を水平方向に連結した特徴量行列である．したがって，観測行列 Y は，文書数 $\big(i = 1, \ldots, N\big)$，特徴量の次元数 $\big(j = 1, \ldots, K\big)$ から構成される N 行 K 列の長方行列である．ここで，次元数 K は 2,071 である．NMF では，観測行列 Y の次元数 K よりも，基底数 M を小さく設定することで，特徴量変換が行える．基底行列 H の基底数を $\big(m = 1, \ldots, M\big)$ とした際の，NMF による観測行列の分解を式 (3) に示す．

$$\big(y_{i,j}\big)_{NK} \simeq \Sigma_{m=1}^{M} h_{j,m} u_{m,i} \tag{3}$$

式 (3) の $(y_{i,j})_{NK}$ は観測行列 Y を表す．また，$h_{i,m}$ は基底行列 H の成分，$u_{m,i}$ は係数行列 U の成分を表す．

行列分解では行列を 2 つの行列の積に分解する．ここで，分解される行列は，一意に決まらない．よって，NMF では，$Y \simeq HU$ の誤差が最小となる，乖離度の最適化が必要である．パラメータ β で統一的に記述した乖離度基準を式 (4) を示す．

$$D_\beta\big(y|x\big) = y\frac{y^{\beta-1} - x^{\beta-1}}{\beta - 1} - \frac{y_\beta - x_\beta}{\beta} \tag{4}$$

式 (4) の β が，$(\beta \to 0)$ のとき Itakura-Saito ダイバージェンス，$(\beta \to 1)$ のとき一般化 Kullback-Leibler ダイバージェンス，$(\beta = 2)$ のとき二乗誤差である [61]．乖離度基準は観測行列 Y の生成プロセスで異なる．本研究では，生成プロセスに，非負の整数の確率分布である Poisson 分布を仮定した．よって，一般化 Kullback-Leibler ダイバージェンスを採用した．

4.4　基 底 選 択

提案手法である基底選択を説明する．提案手法は，NMF の行列分解の結果で，重要な基底の選択する方法である．NMF では，$M < min(K, N)$ のとき，観測行列 Y は，低ランク行列の積で近似することに相当する．よって，観測行列 Y は，基底行列 H と係数行列 U の線形結合で表現できる．線形結合の例を式 (5) に示す．

$$h_{j,1}\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{m,i} \end{pmatrix} + h_{j,2}\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{m,i} \end{pmatrix} \cdots h_{j,m}\begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{m,i} \end{pmatrix} \tag{5}$$

式 (5) の係数行列 U の成分 $u_{m,i}$ の値は文書 i の基底 m への重みを表す．また，観測行列 Y は，文書数 $\big(i = 1, \ldots, N\big)$，特徴量の次元数 $\big(j = 1, \ldots, K\big)$ から構成される N 行 K 列の長方行列である．ここで，特性の異なる 2 種類のテキスト情報の一方のみ，基底の重みが高い場合，特性を表す基底であると解釈できる．式 (5) の文書数 N は，2 種類のテキスト情報が混在している．よって，提案手法では，行列分解後に，行列分解前の観測行列 Y に基づき，係数行列 U を 2 つの係数行列集合に分割する．そして，各基底の重みである係数行列 U の成分 $u_{m,i}$ より，係数行列集合の基底 m の平均値を算出する．2 つの係数行列集合を比較し，基底 m の平均値の差が最も大きい基底は，係数行列集合の特徴を表す基底であると判断できる．したがって，NMF の係数行列 U の係数値で，評価を行いたいテキスト集合の基底選択に応用できる．

4.5　特徴量の評価

提案手法である特徴量の評価を説明する．基底行列 H は，観測行列 Y の特徴量の共起成分がグルーピングされた結果である．NMF の簡略化した分解表現を式 (6) に示す．m は特徴量変換を行う際に指定した次元数，j は観測行列 Y の特徴量の次元数 $\big(j = 1, \ldots, K\big)$ である．ここで，K は 2,071 である．

$$Y \simeq HU \qquad (H = h_{j,1}, \ldots, h_{j,m}) \tag{6}$$

式 (6) の各基底の成分 $h_{j,m}$ の値は，基底 m への特徴量 j の寄与率と解釈できる．よって，重要度の判断基準に採用できる．

上述の節 4.2，節 4.3，節 4.4，節 4.5 のとおり，提案手法は，テキスト情報より抽出した特徴量 (節 4.2) を既存アルゴリズムの NMF で行列分解し (節 4.3)，分解結果の係数行列 U の値で重要な基底行列 H を選択 (節 4.4)，選択した基底行列 H の各特徴量の寄与率を評価 (節 4.5) する特徴量の評価手法である．

5. 評価実験

5.1 実験環境

評価を行うテキスト情報には，2009 年から 2016 年までの 8 年分の環境・循環型社会・生物多様性白書，こども環境白書を用いる．環境白書は PDF (注29) (注30) で提供されている．本研究では，PDF よりテキスト情報を抽出した (注31)．一部こども環境白書（2009 年，2011 年，2014 年）に関しては，2009 年は WEB サービス (注32)，2011 年，2014 年は手作業で抽出した．

提案手法の実装は，プログラム言語 Python (注33) を用いた．単語分割，品詞判定は，MeCab [45] を用いた．LSI および LDA アルゴリズムの実装は，Gensim [62] を用いた．観測行列 Y 作成は，Pandas および NumPy を用いた．NMF および正規化・正則化処理の実装は，scikit-learn (注34) (注35) を用いた．テキスト情報の前処理は，文字コードを UTF-8 に変換，空白記号・改行コードを除去，テキスト情報を「。」「？」「！」で改行，バイト列の欠損行は対象外とした．NMF の文書数は，N=51689 である．テキスト情報の等価な文字は Normalization Form KC (NFKC) (注36) で正規化した．また，形態素解析では，活用形を標準形に変換し，1 文字単語などを削除する処理を実施した．

5.2 結　　果

提案手法を適用し，得られた係数行列 U の値を表 4 に示す．NMF の次元数 M には，M=10 を設定した．重み付け係数値および重み付け係数値の差は各文書集合の平均値である．

表 4 の結果，2 つの文書集合で基底 1 の係数値が最も大きい結果である．よって，係数値の大きさのみでは，基底の判断はできないと判断できる．対して，係数値の差異では，非負値は基底 4，負値は基底 3 という異なる結果である．提案手法では，係数行列集合の基底 m の平均値の差が判断基準である．よって，基底 4 は読みやすさの基底，基底 3 は読み辛さの基底であり，提案手法の判断基準が妥当であると解釈できる．基底 4 の各特徴量の寄与率上位 10 個の特徴量を表 5 に示す．

表 5 より,「文末表現 (です)」が最も寄与率が高い特徴量である．対して，基底 3 の特徴量評価では，寄与率上位に,「文末表現 (た)」が含まれていた．よって，読みやすさは，文末表現が重要であることが明らかになった．また,「単語感情極性対応表 (ポジティブ単語比率)」も重要な特徴量であることが明らかになった．機能表現タグは寄与率上位に複数確認できた．よって，重要な特徴量の種類であると判断できる．一方で,「文字種特徴量 (比率 - ひらがな)」は，寄与率上位だが，基底 3 の特徴量評価の結果，読み辛い文書でも寄与率上位であった．よって，重要な特徴量だが，特性を表す特徴量ではないと判断できる．

表 4　各基底への重み付け係数の平均値

基底	読みやすい文書集合	読み辛い文書集合	係数値の差
1	0.015293	0.015166	0.000126
2	0.003190	0.004047	-0.000857
3	0.000649	0.002243	-0.001594
4	0.001504	0.000740	0.000764
5	0.000073	0.001291	-0.001219
6	0.000157	0.001068	-0.000911
7	0.001347	0.000892	0.000454
8	0.000076	0.000697	-0.000621
9	0.000452	0.000655	-0.000203
10	0.000641	0.000826	-0.000185

表 5　読みやすさの基底 4

特徴量名	特徴量の寄与率
回答文末表現 (です)	0.400483
質問文末表現 (です)	0.400483
機能表現タグ (I-容易)	0.225670
機能表現タグ (I-判断)	0.194876
文字種特徴量 (比率 - ひらがな)	0.066372
機能表現タグ (B-順接限定)	0.029475
機能表現タグ (I-発継続)	0.015669
分類項目一覧表 意味分類コード (1.133)	0.015254
機能表現タグ (I-理由)	0.014842
単語感情極性対応表 (ポジティブ単語比率)	0.013948

6. ま と め

本研究では，特徴量の評価手法を提案した．提案手法では，特徴量行列を NMF で行列分解し，分解結果の係数行列 U の値で重要な基底行列 H を選択し，選択した基底で各特徴量を評価した．評価では，10 種類，31 個，2,071 次元の特徴量を評価した．結果，読みやすい文書では，文末表現，ポジティブ単語比率が重要であった．また，機能表現タグは重要な種類の特徴量であった．今後の課題は，結果の特徴量の考察や概念辞書 (注37) との組み合わせ，NMF のパラメーター β の最適化を行いたい．また，記号表現を物理的に運搬するチャネルや記号の評価箇所を包含して評価したい．そして，特徴量の重み付け係数線形和や可視化手法 [66] に応用し, 読みやすさを選択できる情報推薦や，コンシューマー・インサイト [67] の分析などに活用したい．

7. 謝　　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです．

(注29)：ISO32000-2:2017：https://www.iso.org/standard/63534.html
(注30)：RFC8118：https://www.rfc-editor.org/info/rfc8118
(注31)：PDF ファイルからテキストおよび画像を抽出する方法（Acrobat DC）https://helpx.adobe.com/jp/acrobat/kb/cq06200852.html
(注32)：PDF TXT 変換　PDF を Text に：http://pdftotext.com/ja/
(注33)：Welcome to Python.org：https://www.python.org
(注34)：scikit-learn：http://scikit-learn.org/stable/
(注35)：scikit-learn：https://github.com/scikit-learn
(注36)：Unicode Technical Reports：https://unicode.org/reports/
(注37)：WordNet:http://compling.hss.ntu.edu.sg/wnja/index.ja.html

文　　献

[1] Gantz, John, and David Reinsel. The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East . IDC, 2012.

[2] SIPS ～来るべきソーシャルメディア時代の新しい生活者消費行動モデル概念～，佐藤 尚之，金田 育子，京井 良彦，信澤 宏至，茂呂 譲治，橋口 幸生，宮林 隆吉，電通モダン・コミュニケーション・ラボ，http://www.dentsu.co.jp/sips/index.html（最終閲覧日:2017-08-03）.

[3] 清水 裕士，小杉 考司，対人行動の適切性判断と社会規範―「社会関係の論理学」の構築―，実験社会心理学研究，2010，49，2，132-148

[4] 小林 洋子，日米対人コミュニケーション比較：ビジネスマンの場合，異文化コミュニケーション研究，愛知淑徳大学大学院コミュニケーション研究科異文化コミュニケーション専攻・言語文学研究所，1998-02，1，61-82

[5] 白石 哲郎，ふたつの記号理論と文化の社会学に関する試論 I，佛大社会学，佛教大学，2012-03-25，36，1-14，

[6] 須山 聡，風景印のリテラシー，駒澤地理，駒澤大学文学部地理学教室・駒澤大学総合教育研究部自然科学部門，2012-03，48，15-34

[7] 神嶌 敏弘，推薦システムのアルゴリズム (1)，人工知能学会誌，一般社団法人 人工知能学会，2007-11-01，22，6，826-837，

[8] 小谷 光正，環境マーケティングの進展とグリーンコンシューマーリズム，名古屋学院大学論集. 社会科学篇，名古屋学院大学総合研究所，2016，53，1，13-24

[9] 峰岸 真琴，脳科学は「文法」のありかを特定できるか?; 一般言語学の立場から (第 9 回認知神経科学会)，認知神経科学，認知神経科学会，2005-03，7，1，85-93，

[10] 奥村 学，難波 英嗣：テキスト自動要約 知の科学，オーム社，2005

[11] Barzilay, R.; Elhadad, N., Inferring Strategies for Sentence Ordering in Multidocument News Summarization, Journal Of Artificial Intelligence Research, Volume 17, pages 35-55, 2002.

[12] 大里 彩乃，畳語の研究，言語文化研究，東京女子大学言語文化研究会，2014-03，22，1-16，

[13] 高橋 徹，社会システム分化とゼマンティク：ルーマンにおける社会変動論の一視角，社会学評論，日本社会学会，1999-03-30，49，4，620-634，

[14] 光富 健一，デューイ十進分類法 (DDC)(特集 分類について考える)，情報の科学と技術，一般社団法人 情報科学技術協会，1989，39，11，478-483

[15] 飯田 健夫，感覚情報処理の解明とその社会的貢献，計測と制御，公益社団法人 計測自動制御学会，2002-10-10，41，10，692-695

[16] 大山 正，感覚・知覚測定法 (I)，人間工学，一般社団法人 日本人間工学会，1968，4，1，37-47

[17] 今田 高俊，会長講演 新方法序説に向けて：複雑系、ポストモダンそして自己組織性の視点から，理論と方法，数理社会学会，2003，18，1，1-11

[18] 清川 英男，リーダビリティ研究の概観，淑徳大学研究紀要，淑徳大学，1978-03-20，12，65-82

[19] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.

[20] 柴崎 秀子，リーダビリティー研究と「やさしい日本語」，日本語教育，公益社団法人 日本語教育学会，2014，158，49-65

[21] 難波 英嗣，奥村 学，書き換えによる抄録の読みやすさの向上，情報処理学会研究報告自然言語処理（NL），一般社団法人情報処理学会，1999-09-10，1999，73，53-60

[22] 中山 祐輝，南保 英孝，木村 春彦，レビュー情報を用いた学術本の難易度推定，人工知能学会論文誌，一般社団法人 人工知能学会，2012，27，3，213-222

[23] 家辺 勝文，インターネットの技術基盤の国際化と日本語文書処理 (特集 デジタル時代の日本語)，情報の科学と技術，一般社団法人 情報科学技術協会，2014，64，11，450-455

[24] 福田 誠，吉田 武尚，吉田 誠，樫岡 健史，中学校技術・家庭科教科書 電気領域の表記・表現について，日本教科教育学会誌，日本教科教育学会，2000，23，3，37-42，

[25] 西川 勇佑，中村 雅子，LINE コミュニケーションの特性の分析，東京都市大学横浜キャンパス情報メディアジャーナル，東京都市大学環境情報学部情報メディアジャーナル編集委員会，2015-04，16，49-59

[26] 金 明哲，文節パターンに基づいた文章の書き手の識別，行動計量学，日本行動計量学会，2013-03-28，40，1，17-28

[27] 石田 栄美，安形 輝，野末 道子，文体からみた学術的文献の特徴分析，三田図書館・情報学会研究大会発表論文集，三田図書館・情報学会，2004，33-36

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

[29] Landauer, T. K. and Dumais, S. T., A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, Psychological Review, 104: 2, pp. 211240, 1997.

[30] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022.

[31] 高村 大也，松本 裕治，SVM を用いた文書分類と構成的帰納学習法，情報処理学会論文誌データベース（TOD），一般社団法人 情報処理学会，2003-03-15，44，3，1-10

[32] 奥村 学，マイクロブログマイニングの現在，電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション，一般社団法人電子情報通信学会，2012-01-26，111，427，19-24

[33] 谷口永里子，高橋真理子，最新の女性ファッション雑誌における日本語の特徴の量的分析 -年代差に焦点をあてて；言語処理学会第 22 回年次大会講演論文集，一般社団法人 言語処理学会，2016，D5-1

[34] 佐藤 政光，日本語学習者の語彙習得に関する調査研究 (1) 基本語彙の問題点について，明治大学人文科学研究所紀要，明治大学人文科学研究所，1999-02，44，169-180

[35] 金庭 久美子，専門用語指導のための選定の試み：ニュース語彙を例として，日本語教育方法研究会誌，日本語教育方法研究会，2010-03-27，17，1，2-3

[36] 国立国語研究所，日本語教育のための基本語彙調査，秀英出版，1984

[37] 河合 敦夫，意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，一般社団法人 情報処理学会，1992-09-15，33，9，1114-1122

[38] 松吉 俊，江口 萌，佐尾 ちとせ，村上 浩司，乾 健太郎，松本 裕治，テキスト情報分析のための判断情報アノテーション，電子情報通信学会論文誌. D 情報・システム，一般社団法人 電子情報通信学会，2010-06-01，93，6，705-713

[39] 上岡 裕大，成田 和弥，水野 淳太，乾 健太郎，述部機能表現に対する意味ラベル付与，研究報告音声言語情報処理（SLP），一般社団法人 情報処理学会，2014-05-15，2014，9，1-9

[40] 福田 一雄，日本語モダリティ覚え書き (その 1)，言語の普遍性と個別性，新潟大学大学院現代社会文化研究科「言語の普遍性と個別性」プロジェクト，2014-03，5，1-13

[41] 西原 陽子，松村 真宏，谷内田 正彦，Q&A コミュニティでの質疑応答パターンの理解，人工知能学会全国大会論文集，一般社団法人 人工知能学会，2008，22，1-4

[42] 横山 友也，宝珍 輝尚，野宮 浩揮，佐藤 哲司，文章の特徴量を用いた質問回答文の印象の因子得点の推定，日本感性工学会論文誌，日本感性工学会，2013，12，1，15-24，

[43] 細貝 亮 メディアが内閣支持に与える影響力とその時間的変化: 新聞社説の内容分析を媒介にして，マス・コミュニケーション研究，日本マス・コミュニケーション学会，2010，77，225-242，

[44] 中尾 桂子，品詞構成率に基づくテキスト分析の可能性：メール自己紹介文、小説、作文、名大コーパスの比較から，大妻女子大学紀要. 文系，大妻女子大学，2010-03，42，128-101

[45] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, In Proceedings of the Conference on Empirical

Methods in Natural Language Processing (EMNLP ’ 04). 230–237.
[46] 関根 聡，竹内 康介，拡張固有表現オントロジー，言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」，一般社団法人 言語処理学会，pp. 2326，2007
[47] 藤井 裕也，飯田 龍，徳永 健伸，Wikipedia 記事を利用した曖昧性のある表現の固有表現クラス分類，言語処理学会第 16 回年次大会，一般社団法人 言語処理学会，pp.1518，2010
[48] 乾 孝司，奥村 学，テキストを対象とした評価情報の分析に関する研究動向，自然言語処理，一般社団法人 言語処理学会，2006-07-10，13，3，201-241
[49] 小林 のぞみ，乾 健太郎，松本 裕治，立石 健二，福島 俊一，意見抽出のための評価表現の収集，自然言語処理，一般社団法人 言語処理学会，2005，12，3，203-222
[50] 東山 昌彦，乾 健太郎，松本 裕治，述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，一般社団法人 言語処理学会，2008，584-587
[51] Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL ’05). Association for Computational Linguistics, Stroudsburg, PA, USA, 133-140.
[52] 小林 のぞみ，乾 健太郎，松本 裕治，意見情報の抽出／構造化のタスク仕様に関する考察，情報処理学会研究報告自然言語処理（NL），一般社団法人 情報処理学会，2006-01-13，2006，1，111-118
[53] 岡崎 直観，Web 文書からの人の安全・危険に関わる情報の抽出，言語処理学会第 18 回年次大会発表論文集，一般社団法人 言語処理学会，2012，895-898
[54] 小谷 龍ノ介，新聞記事における偏向性の定量評価，法政大学大学院紀要 理工学・工学研究科編，法政大学大学院理工学研究科，2017-03-31，58
[55] 徳久 良子，乾 健太郎，松本 裕治，Web から獲得した感情生起要因コーパスに基づく感情推定，情報処理学会論文誌，一般社団法人 情報処理学会，2009-04-15，50，4，1365-1374
[56] 丹治 広樹，村田 真樹，柿澤 康範，Stijn，De Saeger，鳥澤 健太 トラブルを表す文の Web からの抽出，言語処理学会第 15 回年次大会発表論文集，一般社団法人 言語処理学会，2009，140-143
[57] 野宮 浩揮，宝珍 輝尚，顔特徴量の有用性推定に基づく特徴抽出による表情認識，知能と情報：日本知能情報ファジィ学会誌，日本知能情報ファジィ学会，2011-04-15，23，2，170-185
[58] 永田 賢二，岡田 真人，スパースモデリングを用いた特徴選択と地球科学データ解析 (特集 スパースモデリング: 情報処理の新しい流れ)，応用数理，一般社団法人 日本応用数理学会，2015，25，1，5-9
[59] 小林 重信，吉田 幸司，山村 雅幸，GA によるパレート最適な決定木集合の生成，人工知能学会誌，一般社団法人 人工知能学会，1996-09-01，11，5，778-785
[60] 平 博順，春野 雅彦，Support Vector Machine によるテキスト分類における属性選択，情報処理学会論文誌，一般社団法人情報処理学会，2000-04-15，41，4，1113-1123
[61] 亀岡 弘和，非負値行列因子分解，計測と制御，公益社団法人 計測自動制御学会，2012-09-10，51，9，835-844
[62] Rehruvrek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (p./pp. 45–50), May, Valletta, Malta: ELRA.
[63] N. Halko, P. G. Martinsson, and J. A. Tropp. 2011. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Rev. 53, 2 (May 2011), 217-288.
[64] Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for Latent Dirichlet Allocation. In Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (NIPS’10), J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), Vol. 1. Curran Associates Inc., USA, 856-864.
[65] Roger B. Bradford. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM ’08). ACM, New York, NY, USA, 153-162.
[66] 須鎗 弘樹，ベイジアンネットワーク入門 (1)，日本医用画像工学会，2003-09-25，21，4，315-318
[67] 亀井昭宏：電通広告事典，電通，2008

付　　録

本研究で特徴量抽出に用いた既存研究の辞書を表 A・1 に示す．単語感情極性対応表は，日本語の辞書を用いた．また，見出し語には，辞書の見出し語と読みを用いた．NAIST-JENE は，Wikipedia の見出し語に対し，関根の拡張固有表現階層の定義に基づき固有表現クラスを付与した辞書である．日本語教育基本語彙，意味分類体語彙表，分類項目一覧表は調査研究目的に作成された一覧表である．本研究では下記の基準で，加工・見出し語の選定した．

（1） 見出し語の加工
　　空白記号,「-」「0」「など」を除去
（2） 対象外の見出し語
　　「」「-」「〔」「→」「/」「(」「」「その他」を含む語

表 A·1　既存研究の辞書一覧

種別	特徴量名	文献/辞書
表層情報	文字種	Unicode 10.0(注a)
語種	語種	意味分類体語彙表 (注b)
基本語	基本語 (1)	日本語教育基本語彙 (注b)
基本語	基本語 (2)	意味分類体語彙表 (注b)
基本語	基本語 (3)	分類項目一覧表 (注b)
意味属性	意味分類コード (1)	日本語教育基本語彙 (注b)
意味属性	意味分類コード (2)	意味分類体語彙表 (注b)
意味属性	意味分類コード (3)	分類項目一覧表 (注b)
言語表現	機能表現	機能表現タグ付与コーパス [38](注c)
文末表現	文末モダリティ	文献 [40]
文末表現	質問文末表現	文献 [41]
品詞	IPA 品詞	ipadic version(注d)
固有表現	固有名詞	ipadic version(注d)
固有表現	拡張固有表現	NAIST-JENE(注e)
評価表現	評価極性情報	
	（用言編）	日本語評価極性辞書 [49](注c)
	（名詞編）	日本語評価極性辞書 [50](注c)
評価表現	感情極性	単語感情極性対応表 [51](注f)
評価表現	評価値表現	評価値表現辞書 [52](注g)

（注a）：Unicode 10.0 Character Code Charts
http://www.unicode.org/charts/
（注b）：『日本語教育のための基本語彙調査』データ
http://mmsrv.ninjal.ac.jp/bvjsl84/
（注c）：機能表現タグ付与コーパス，日本語評価極性辞書
http://www.cl.ecei.tohoku.ac.jp/index.php
（注d）：ipadic version 2.7.0 ユーザーズマニュアル
http://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf
（注e）：NAIST Japanese ENE Dictionary on Wikipedia
https://github.com/masayu-a/NAIST-JENE
（注f）：単語感情極性対応表
http://www.lr.pi.titech.ac.jp/t̃akamura/pndic_ja.html
（注g）：評価値表現辞書
http://www.syncha.org/evaluative_expressions.html