視覚情報を活用したテキストマイニング辞書構築システムの提案

山下 紗苗† 小比田 涼介‡ 上 泰†

†明石工業高等専門学校 機械・電子システム工学専攻
‡日本アイ・ビー・エム株式会社 東京基礎研究所
†me1912@s.akashi.ac.jp ‡kohi@ibm.com †kami@akashi.ac.jp

1 はじめに

本稿では，視覚情報を用いて効率的にテキストマイニング辞書を構築するシステムを提案する．テキストマイニング辞書とは，何らかの観点で意味を共有する単語の集合であり，分析の観点を表す．しかし辞書はドメインや分析目的に依存するため，分析ごとにその都度作り直さねばならない．したがって，その構築を支援するシステムが求められている．

辞書構築の支援のひとつに，図 1 のようにシステムがユーザの求める単語を推薦する方法がある [1]．この方法では，ユーザが辞書に登録したい単語を指定すると，システムは図 2 のように候補となる単語のリストを提示する．単語に対して辞書に登録するかどうかのアノテーションを行うと次の候補単語が提示され，この繰り返しによって辞書を拡充していく．このような**リストベース**の辞書構築支援は，あらかじめ分析の観点が定まっている，つまり，作りたい辞書が決まっている際に効果を発揮し，膨大な単語群から辞書に登録したい単語を探す手間を削減する．

しかし分析の初期段階などでは，様々な辞書を構築していく中で分析の観点を見出すことがよくある．我々はそのような場面を**探索的辞書構築**と呼ぶ．探索的辞書構築には，リストベースの方法は必ずしも適しているとは言えない．なぜならリストベースでは，ユーザは初期段階から作りたい辞書の概要を想像することが求められ，また，探索の幅も提示される候補に狭められてしまうからである．テキストマイニングは，情報源に潜在する興味深いパターンを試行錯誤しながら見つけ出し，有用な情報を得ることを目指している [2] ため，探索的辞書構築の支援は重要な課題である．

探索的な辞書構築の支援システムについて，次の 3 つの要件が考えられる．

(i) **候補単語全体を俯瞰できる**
(ii) **探索深度を調節できる**
(iii) **繰り返し行う作業を効率化できる**

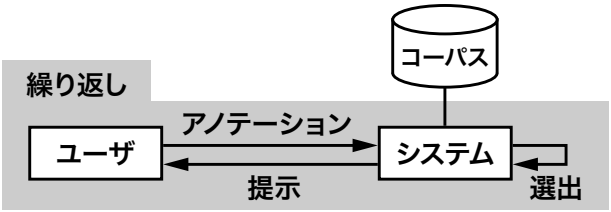


図 1: 小比田ら [1] による対話的な辞書構築の流れ

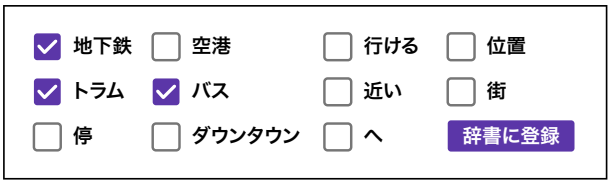


図 2: 候補単語の表示例（リストベース）

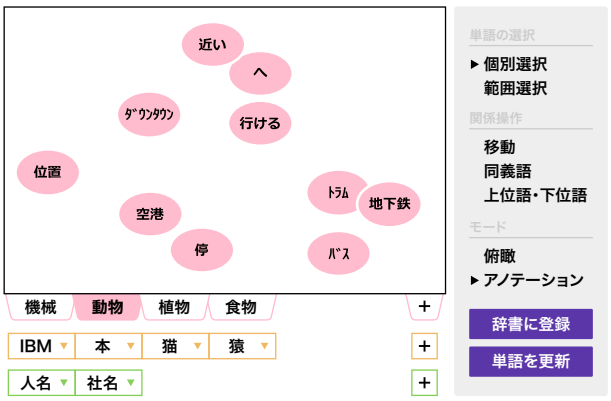


図 3: 提案システムの画面（マップベース）

探索的な辞書構築の支援を目的としたときに，候補単語の提示において重要なのは，その仕組が**新たな発見を促す**ように設計されていることである．すなわちシステムは，ユーザがまだ気づいていない分析の観点に気づくように候補単語を提示できればよい．そのためには，提示される候補は豊富でありながらも見やすさが担保されており，ユーザが単語全体を俯瞰できるような形が望ましい (i)．さらに探索には，全体を俯瞰する，大まかなあたりをつける，詳しく見る，などの段

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

階があるため，これを調節できる必要もある (ii)．また，辞書構築は試行錯誤の繰り返しであるため，見た目や操作がユーザの直感に沿うように考えられていると，より効率的な辞書構築につながる (iii)．

我々は，人間の視覚特性に基づいたマップベースの辞書構築支援システムを提案する．図 3 のように単語を地図的に表示することで，単語全体の俯瞰や意味関係の素早い把握を可能にする．加えて，テキストマイニング辞書の特性や人間の直感に符合する機能・操作によって，辞書構築における試行錯誤を効率化する．

2　視覚に関する人間の特性

人間の特性を考慮することで，より使いやすいシステムを構築できる．この章では近接の要因とプライミング効果について述べる．

本システムでは，ユーザが新しい分析の観点に気づけるように候補単語を提示したい．大量の単語の中から潜在的な共通点を見つけやすくするためには，近接の要因を用いて視覚的に単語のグループを作ればよい．また効率化の観点から，単語を辞書に登録するかどうかの判断は頻繁に発生するため，迅速に行いたい．似た意味の単語が近くに配置されるようにすれば，プライミング効果により判断速度が上がると期待できる．

2.1　ゲシュタルト法則の近接の要因

ゲシュタルト法則は，人間が視覚を使ってどのように事物を見分け，とらえているのかについて説明するものである [3]．この法則には近接，類同，連続，閉合，対称性などの要因があるが，本稿では近接の要因に注目する．近接の要因 [4] は，表示されているオブジェクトどうしの相対的な視覚距離が，オブジェクトのグループ化に影響を与えるというものである．他のオブジェクトよりも視覚距離の近いオブジェクトどうしは同じグループに属するように見えるが，離れたオブジェクトどうしは同じグループには見えない [3]．

広川ら [5] は，多くの UI 要素を見やすくまとめる手段としてこれを紹介している．例えば，iPhone のホーム画面におけるアイコン群はドック部と上部に分かれているが，境界にスペースがあり近接の要因がはたらくため，別々のグループだと認識できる．

2.2　プライミング効果

プライミングとは，先行刺激（プライマー）が後続刺激（ターゲット）の処理に影響を及ぼすことをいう．プライマーとターゲットの間に意味的関連がある場合，ターゲットの理解が促進される [6]．プライマーとターゲットが同一の場合のプライミングを直接プライミング，プライマーとターゲットが何らかの関係がある場合のそれを間接プライミングという [7]．内海 [8] は間接プライミングについて，3 つの単語「dog」「cat」「pen」を用いて次のような例を挙げている．dog に続いて提示される cat に対する反応（単語か非単語かを判断させる語彙性判断課題や，単語を読み上げさせる命名課題など）は，pen に続いて提示される cat に対する反応よりも反応時間が短い．つまり，dog と cat は意味的に類似しているが，pen と cat は類似していない，といった単語間の意味的な類似性が単語認知に影響を及ぼすのである．本稿では，この間接プライミングのことを単にプライミングと呼ぶ．

猪原ら [9] によれば，プライミング効果を引き起こしうる概念間関連には数多くの種類があることが分かっている．猪原らは概念間関連のひとつとして潜在意味解析による単語間類似度を挙げ，語彙性判断課題と命名課題において，この単語間類似度とプライミング効果の間には正の相関関係があることを示した．

3　辞書構築システムの提案

我々は，2 章のような視覚特性を活用したマップベースのテキストマイニング辞書構築システム（図 3）を提案する．このシステムは Web アプリケーションとしてブラウザ上で動作し，ユーザのアノテーションとサーバによる候補単語の選出を繰り返すことで，対話的に辞書を構築できる．Word2Vec などを用いて単語の意味表現を 2 次元空間に写像し，その結果を当該単語の位置として取り扱うことで，単語間の類似度が高いほどマップ上で近くに配置されるようにする．

このシステムは主に以下の 4 つの機能を有する．

マップ表示　候補単語間の関係を 2 次元上の位置関係で表し，マップとして画面に表示する

辞書登録　候補単語を選択し，辞書に登録する．候補単語は適宜更新できる

関係操作　マップ上で候補単語の位置を移動することで，意味空間上の位置も移動させる．候補単語どうしを，同義語や上位語･下位語などで関係づける

エントリ管理　辞書，同義語，上位語・下位語の構造を管理する

マップ表示を俯瞰モード，その他の 3 つをまとめたアノテーションモードとも呼ぶ．続いて，これらの機

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

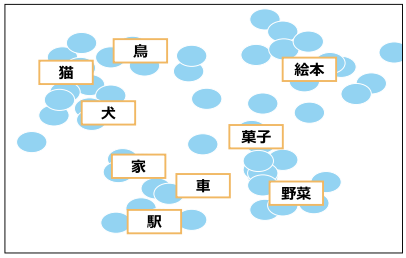
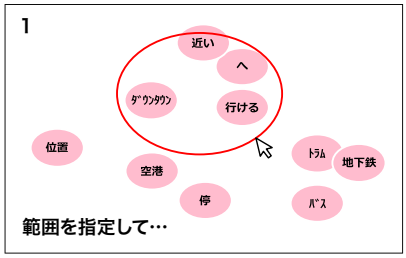
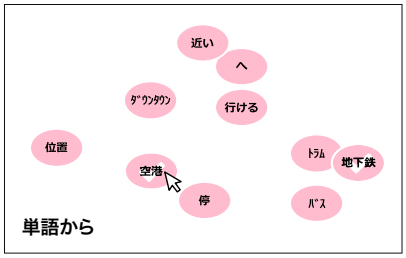


図 4: 俯瞰モード





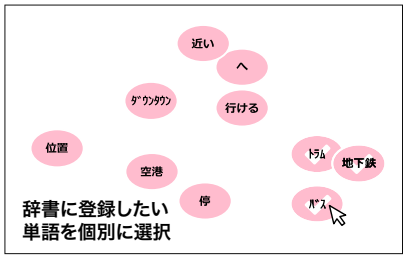


図 5: 辞書登録（個別選択）

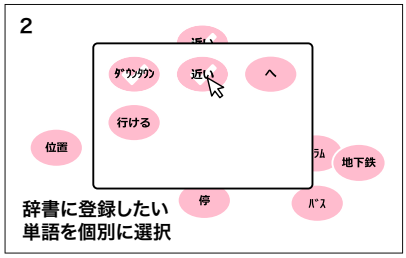


図 6: 辞書登録（範囲選択）

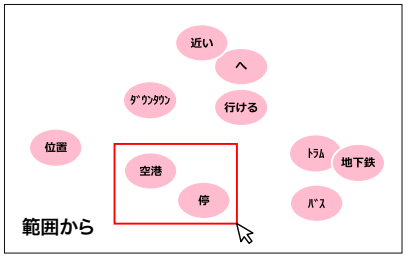
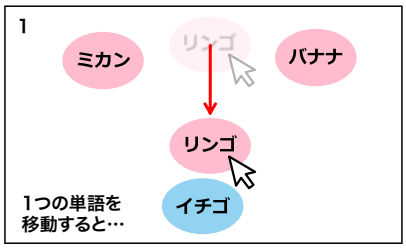


図 7: 候補単語を更新する



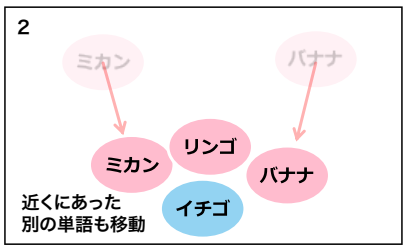
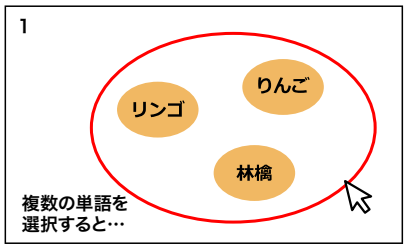


図 8: 移動



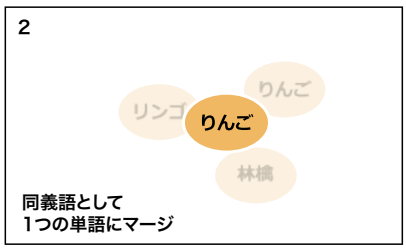
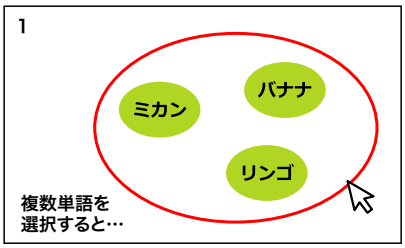


図 9: 同義語



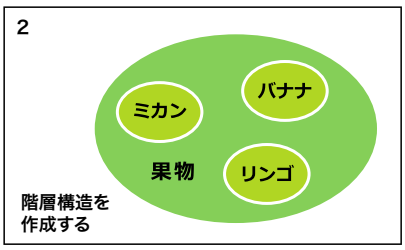


図 10: 上位語・下位語

能がどのように探索的な辞書構築支援システムの要件 (i)〜(iii) を満たしているか，またどのようにユーザに新たな発見を促すかを説明する．

3.1 マップ表示

渡辺はテキストマイニングのための可視化技術について，情報の外観や情報間の関連性を提示することの有用性を 2 つ述べている [10]．ひとつは，探している情報を見つけだすためのガイドになること，そしてもうひとつは，主要なテーマは何か，テーマ間の関連はどうなっているのかといった傾向を把握するための支援になることである．本システムのように探索的な辞書構築を目的とする場合には，後者が関係する．

俯瞰モードは図 4 のような画面をもつ．このモードでは候補単語全体を俯瞰して探索できる (i) ほか，ズーム操作によって探索深度を調節できる (ii)．候補単語の分布を把握することが目的であるため，単語名は周辺の単語群を代表して数個だけを表示する．2.1 節で述べた近接の要因により，近い位置にある単語どうしは同じグループに属しているように見えるため，グループどうしの関連や傾向から新しい分析の観点を見つけるための支援ができる．また，類似度が高い単語はマップ上でも近くに配置されるようになっているため，2.2 節で述べたプライミング効果により，辞書に登録するかどうかの判断が速くなると期待できる (iii)．このように，探索的に辞書を構築したい場合は，単語間の類似度を視覚的に表現しやすいマップベースのほうが，リストベースよりも優れている．

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.

3.2 辞書登録

単語を選択して辞書登録ボタンを押すことで，単語を辞書に登録できる．選択方法には図 5 の個別選択と図 6 の範囲選択がある．範囲選択では一度に複数の単語を選択できるため，辞書構築が効率的になる (iii)．

また，ユーザは適宜，表示する候補単語を更新できる．図 7 のように単語や範囲を指定すると，サーバは，指定された単語やその範囲に含まれる単語をもとに新たな候補単語を返す．

3.3 関係操作

候補単語間の関係操作には，単語の移動，同義語の作成，上位語・下位語の作成がある．単語の移動とは，図 8 のようにマップ上で単語を移動することで，その単語と周辺の単語のベクトル空間上での位置も更新される機能である．これによって単語ベクトル構築時には得られなかった情報を人手で与えることができ，単語間の新たな関係に気づくことができる．

同義語や上位語・下位語は，図 9 と図 10 のような操作で作成する．ユーザは選択した単語の中から代表語をひとつを選び，以降はこの単語だけをマップに表示する．このように，単語に対する移動，結合，階層の操作を単語や辞書が持つ概念と結びつけることで，より直感的なアノテーションができるようになる (iii)．

3.4 エントリ管理

作成した辞書，同義語，上位語・下位語の構造をエントリと呼ぶ．ひとつのコーパスに対して複数の辞書を構築することができ，作成したエントリはコーパスごとに共有される．例を図 11 に示す．これらのエントリはコピーが可能で，例えばより細分化されたエントリを作る際に効率的である (iii)．

図 11: エントリ管理エリア

4 おわりに

本稿では，テキストマイニングにおける辞書の構築方法に注目し，人間の視覚特性を考慮した辞書構築支援システムのインタフェースを提案した．このシステムは探索的な辞書構築の支援を目的とし，そのために，単語全体を俯瞰し，探索深度を調節し，繰り返し発生する作業を効率化できる機能を備えている．この機能を活かすため，我々は近接の要因とプライミング効果を根拠に，単語を類似度に基づいて地図的に提示するマップベースの方法を採用した．

本稿はシステムのインタフェースを提案したものであり，現時点では，候補単語の選出や提示がうまくできることが前提となっている．実際に単語間の類似度を 2 次元地図上の距離として正しく表現できるかは，今後の課題である．

参考文献

[1] 小比田涼介, 那須川哲哉, 吉田一星, 金山博. ドメイン・ユーザー指向辞書の対話的構築. 第 25 回言語処理学会年次大会発表論文集, pp. 1273–1276, 2019.

[2] Ronen Feldman and James Sanger. テキストマイニングハンドブック. 東京電機大学出版局, 2010. 辻井潤一 監訳, IBM 東京基礎研究所テキストマイニングハンドブック翻訳チーム 訳.

[3] Jeff Johnson. UI デザインの心理学－わかりやすさ・使いやすさの法則. 株式会社インプレス, 2016.

[4] M. Wertheimer. Laws of organization in perceptual forms. *A source book of Gestalt psychology*, pp. 71–88, 1938. Ellis, W. 英訳.

[5] 広川美津雄, 井上勝雄, 岩城達也, 加島智子. 直感的インタフェースデザインの設計論の基礎的考察－体制化と親近性の視点からのアプローチー. 日本感性工学会論文誌, Vol. 13, No. 5, pp. 543–554, 2014.

[6] D.E. Meyer and R.W. Schvaneveldt. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, Vol. 90, No. 2, pp. 227–234, 1971.

[7] 太田信夫. 長期記憶におけるプライミングー驚くべき潜在記憶（implicit memory）ー. 心理学評論, Vol. 31, No. 3, pp. 305–322, 1988.

[8] 内海彰. 言語と類似性. 人工知能学会誌, Vol. 17, No. 1, pp. 8–13, 2002.

[9] 猪原敬介, 楠見孝. 潜在意味解析に基づく概念間類似度の心理学的妥当性. 心理学評論, Vol. 54, No. 2, pp. 101–122, 2011.

[10] 渡辺勇. ビジュアルテキストマイニング. 人工知能学会誌, Vol. 16, No. 2, pp. 236–232, 2001.

Copyright(C) 2020 The Association for Natural Language Processing.
All Rights Reserved.