LDA を使った専門用語の教師なしクラスタリング

黒田 航[1]　相良 かおる[2]　東条 佳奈[3]　麻 子軒[4]　西嶋 佑太郎[5]　山崎 誠[6]
[1]杏林大学　[2]奈良先端技術大学院大学　[3]大阪大学　[4]関西大学　[5]京都大学　[6]国立国語研究所

概要

医療，(政治) 経済，法律，出版の 4 分野の用語の，Latent Dirichet Allocation (LDA) を使った教師なしクラスタリングの実用性を検討した．単語を document とし，文字の (不) 連続 n-gram ($n < 4$) を term として LDA を実行した．結果から判った事は 2 つある．小さな topic 数で用語を分野別に分ける課題が実現できる．大きな topic 数では語構成パターンの分類が可能になる．

1　はじめに

数多くの用語が目の前にある時，それらをうまく分類したいという欲求は自然である．分類はすべての情報処理の始まりであり，不可欠なものである．

分類という課題に内在する問題は，2 つある．一つ目は労力の問題，二つ目は精度の問題である．これらは通常，相反する．自動分類は労力はかからないが精度が低く，人手分類は精度はそれなりだが労力がかかる．とは言うものの，人手分類がいつも最良の分類を与えるかどうかは明らかではない．特に人手分類は，観点が多元的な分類[1)]と粒度が可変な分類を実現しない．

『実践医療用語_語構成要素試案表 Ver. 2』[2)] [2] には 7,087 事例が収録されているが，これでも網羅的とは言えない．網羅性を重視するなら，この数倍の分量の用語を集める必要があるだろう．そうして得られた規模の医療用語をすべて人手で分類するという課題は，非現実的である．

本研究は Latent Dirichlet Allocation (LDA) [3, 4] を使って，それなりの精度の用語の自動分類が実現できる事を示す．

1)　例えば {バセドー氏病, 心臓病, 心臓肥大} を分類する時，人は {{バセドー氏病, 心臓病}, 心臓肥大} の様に主要部で分類しがちであるが，これは {バセドー氏病, {心臓病, 心臓肥大}} の分類を排除する．このような複数解の存在を許容する分類は Formal Concept Analysis (FCA) を使わないと実現できない [1].

2)　https://www.gsk.or.jp/catalog/gsk2020-g/

2　解析法

2.1　データ

解析の対象としたデータは，次の 4 種類の用語 M(edical), E(conomic), J(uridical), P(ublishing) の，出現バランスを考慮して混合した部分的にランダムな (文字長が 12 を超えるもの除く) 1,997 語である．M) 『実践医療用語_語構成要素試案表 Ver. 2』 [2] に収録されている 7,087 事例のうち，MLMA の記述 [5] が終わっている 3,530 語，E) 政治経済用語として『基礎研 WEB 政治経済学用語事典』[3)]から抽出した 507 語，J) 法律用語として『法テラス』の「法律や裁判で使うことば【用語集】」[4)]から抽出した 207 語と『法令用語日英標準対訳辞書』[5)]から選別した 442 語の和，P) 出版用語として『製本用語集』[6)]から抽出した 3,834 語からサンプルとして得た 1,529 語．

今回の解析に使ったのは，事例数が最小の E の規模に併せて，M, J, P のそれぞれからランダムに 500 事例を選んで，混合構成した 1,997 事例である．これを以下，データ D と呼ぶ．表 1 に D の見本を示す．D は §2 の解析で共通に使用された．

表 1　D の見本 6 例

用語	医療	経済	法律	出版
審問	0	0	1	0
後頭葉	1	0	0	0
金融政策	0	1	0	0
ISBN	0	0	0	1
前方脱臼	1	0	0	0
婚姻届	0	0	1	0

ここに示した例を見て，注意深い読者は分野分類は排他的にならないと気づかれると思う．これは経済用語と法律用語の間で顕著であるが，他の分野間でも起きない訳ではない．これは用語の分類が本

3)　http://www.kisoken.org/webjiten/kisokenjiten.html
4)　https://www.houterasu.or.jp/forforeignnationals/yougoyasasiinihongo/index.html
5)　https://www.japaneselawtranslation.go.jp/ja/dicts/download
6)　https://sei-hon.jp/glossary/index.html

来，決定的な解を得られない課題である事を意味している．結果の評価ではソフトクラスタリングを使う必要があり，本研究では t-SNE [6] を使った．

2.2 LDA のパラメター

LDA には gensim パッケージ (v4.2.x)[7] の LdaModel を使用した．LDA は制御すべきパラメターの数が多く，異なる値で結果が変わる．それ故に，妥当な結果を得るには試行錯誤が必要だった．その上，現時点で最良だと見なせる結果も真の最良設定である保証はない．

後で詳しく述べるが，topic 数に関していうと，最適値は目的次第という面がある．目的が用語の大分類であれば，topic 数は少ない方が良い結果を与える．その一方，分類の目的が語構成のパターン抽出であれば，topic 数はそれなりに大きな値が良い結果を与える．具体例は後で詳しく述べる．

2.3 DTM 構築のパラメター

まず，LDA の入力となる文書単語行列 (document-term matrix: dtm) 構築の際に，どんな要素を term とするかを決める必要がある．それが決まった後で，出現傾向に応じて選別をかける必要がある．文 (書) を対象にした LDA では，文 (書) を doc(ument) とし，(通常は lemma 化された) 単語を term とする．この設定を素直に用語クラスタリング課題に適用すれば，語が doc で文字 が term となる (“文: 語 = 語: 文字” のアナロジーが成立)．これは素直な設定だが，doc を bag-of-words (bows) としてエンコードすると，doc 中の word の順序はモデルに反映されない．

2.3.1 文字の不連続 n-gram を term に利用

LDA は部分と全体に共通の生成源を推定する手法だと抽象化すると，部分は全体の分割である必要はない．部分間の共起度 d を導入し，要素の 1-gram の場合は $d = 0$，要素の (不) 連続な 2-gram の場合は $d = 1$ とすると，任意の全体について d の次数に応じて部分を定義できる．要素の不連続 n-gram (= k-skip n-gram [7]) がそのような要素である．本調査では，用語の文字 $\{1, 2, 3\}$-gram (これらは doc の $d = 0$ の構成要素) の他に，不連続 (skippy) な文字 2-gram (doc の $d = 1$ の構成要素)，不連続な文字 3-gram (doc の $d = 2$ の構成要素) を調査し，結果を比較した[8]．

不連続 n-gram の生成は計算資源を要求し，モデルの収束を遅くするので，不連続性に上限を設けるのが自然だと考えた．具体的には，用語の文字数が L の時，不連続性 k が $L * 2/3$ より大きくならない条件で，文字の k 不連続 n-gram を生成した．

2.3.2 過剰使用と過少使用への対応

LDA に与える dtm は term の最低頻度と使い過ぎ指標で制御できる．事前調査から，最低頻度は 3，使い過ぎ指標は 0.03 とした[9]．文字の skippy n-gram を使うと term 数が多くなるので，使い過ぎ指標が 0.05 より大きくなると，相当に計算資源を要する．

2.3.3 最適 topic 数の事前調査

LDA の topic 数の最適性はもっとも確定し難いパラメターである．と言うのは，結果的には，topic 数は目的に依存するもので，最適値を確定的に規定するのは意味がない判明したからである．だが，その結論に至った経緯を説明すべきだろう．

まず，R 用の ldatuning パッケージ[10] を使って，term の設定条件 [文字の i, 1-gram, ii. 連続 2-gram, iii. 連続 3-gram, (不) 連続 1-gram, (不) 連続 2-gram, (不) 連続 3-gram)] のそれぞれに最適なトピック数を推定した．ただし，n-gram は (通常と違って) $(n-1)$-gram を包含する事とした．理由は 2 つある: i) こうする事で計算量が増える他に害が生じる訳ではない．ii) こうしないと短い用語 (具体的には 1 文字や 2 文字の用語) に適切なエンコードが得られない．

この手順で得られた最適な topic 数は，term を文字の連続 2, 3-gram とする場合で 15–20 個，不連続 2, 3-gram とする場合で 25–30 個であった[11]．1-gram の最適値は 20 個以内である事しか決まらなかった．

2.4 LDA の結果の評価

以上で設定で LDA を実行した結果は，次の 4 種類の方法で評価した．解析 1a) topic ごとの関連 term ランクの自然さ，解析 1b) pyLDAvis を使った topic と term の対応の良さ，解析 2) 用語 (doc) の tSNE を使ったクラスタリング，解析 3) 用語 (doc) の階層クラスタリング (付録の図 **??**に例を示した)．ただし，

7) https://radimrehurek.com/gensim/
8) 調査で skippy 文字 2-gram で十分な性能が得られると判明したので，$3 < n$ の解析は実行しなかった．
9) 調査に使用した gensim パッケージの dtm 濾過で minfreq = 3 と abuse_threshold = 0.03 とした．
10) https://github.com/nikita-moor/ldatuning
11) これは不連続 n-gram の方が担える情報量が多い事を示唆する．不連続 n-gram が構成要素間の相対順序 (= 原初的な統語構造?) をエンコードしていると考えれば，納得できる．

これらはどれも質的評価であり，定量評価には至っていない．

本発表では紙面の制限に拠り，解析 3 の結果のみを示す．他の結果，並びに Jupyter Notebook 上の実行コードは GitHub [12]で入手可能である．

t-SNE で妥当な結果を与えると思われる perplexity の値を求めるのは，それなりに苦労した．適切な perplexity が LDA の topic 数に依存するのか，事例数に依存するのかが不明だった．事前調査の結果，事例クラスタリングでは，事例数に依存した大きな値で良い結果が得られる場合がある事が判った．ただ，これは topic 数が大きく skippy n-の n が大きい場合に起こる事で，topic 数が小さい場合には，最適な perplexity 値は単純に topic 数に依存するように見えた．様々な値を試して判った事は，topic 数に応じて最適な perplexity が変化するという事である (それ故，当初は perplexity の最適値が単純に topic 数に依存するように見えた)．だが，実際には最適な perplexity 値は topic 数ではなく事例数に依存する．t-SNE の perplexity は今の時点では，LDA の topic 数を 5, 10, 20 などに固定してから，term の型 (文字の (不) 連続 n-gram) ごとの個別の最適値を探している．最適性の評価法を数値化できていないので，判断は主観評価である．

3 解析結果

3.1 topic 数 5 で実行した結果の可視化

topic 数 5 で実行した LDA クラスタリングの結果の t-SNE による可視化を図 1–図 2 に示す．紙面の都合で，1-gram と skippy 2-gram のみである．

3.2 topic 数 10 で実行した結果の可視化

topic 数 10 で実行した LDA クラスタリングの結果の t-SNE による可視化を図 3–図 4 に示す．紙面の都合で，1-gram と skippy 2-gram のみである．

3.3 topic 数 20 で実行した結果の可視化

topic 数 20 で実行した LDA クラスタリングの結果の t-SNE による可視化を図 5–図 6 に示す．紙面の都合で，1-gram と skippy 2-gram のみである．

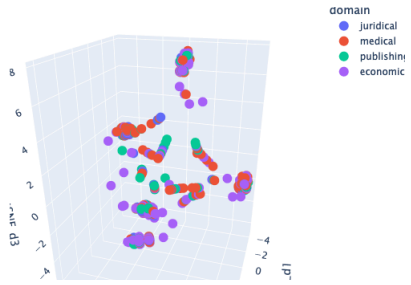

図 1 t-SNE 3D [topic 数: 5; 1-gram; perplexity: 255]

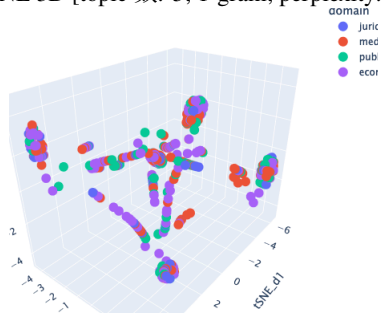

図 2 t-SNE 3D [topic 数: 5; skippy 2-gram; perplexity: 255]

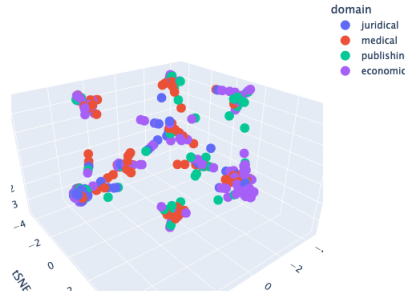

図 3 t-SNE 3D [topic 数: 10; 1-gram; perplexity: 255]

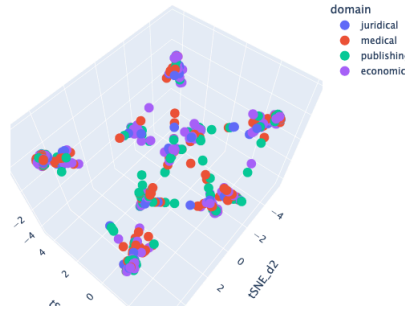

図 4 t-SNE 3D [topic 数: 10; skippy 2-gram; perplexity: 255]

12) https://github.com/kow-k/LDA-mixed-terms

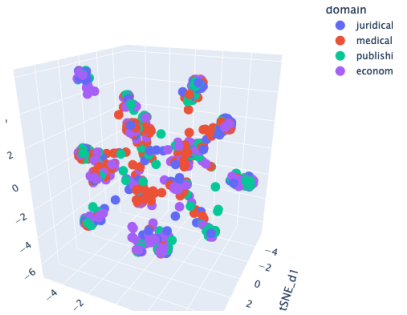

図 5　t-SNE 3D [topic 数: 20; 1-gram; perplexity: 255]

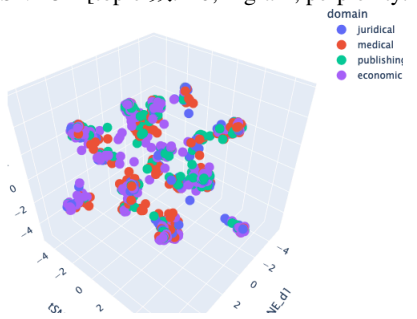

図 6　t-SNE 3D [topic 数: 20; skippy 2-gram; perplexity: 255]

3.4　結果のまとめ

結果を比較してわかるのは，topic 数が違うと解析結果の質が変わる事である．i) topics 数が多いと，大きな値の perplexity でも局所構造を保持しミクロ構造が捉えられるが，マクロ構造は不明瞭になる．i) topic 数が少ない場合，perplexity が大きくなるとデータ点が凝集し局所構造が潰れるが，マクロ構造を捉えやすくなる．用語の分野分類が目的であれば，表現力を失わない範囲でなるべく少ない topic 数で LDA を実施すれば良いと言える．

分野分類が目的でないなら，topic 数を大きくする事で語形成パターンの抽出が可能になるが，事例数が多い場合，perplexity 値の最適化に手間がかかる．

4　考察と結論

4.1　研究の流れ

この研究は LDA を用語 (= 単語) を document とし，文字を term とする設定で使えば，教師なし条件で医学用語がうまく分類できるのではないか？という想定で始められた[13]．最初は，文字の連続した {1, 2, 3}-gram を term に使った．それなりの結果が得られたものの，評価が難しかった．グループ化はうまく実現されているように見えるものの，それが最適なのか判定が難しかった．この段階で，要素の相対的順序が捉えられていない事が結果に影響している事が疑われた．具体的に言うと，「症」や「病」で終わっている用語がクラスターをなしていなかった．これに対処するため，term に不連続な *n*-gram を使って見て，それなりに満足できる結果が得られた．

ただ，この改良は topic 数の最適化をすぐに実現した訳ではない．LDA が実現するのは教師なしのクラスタリングなので，「正解」との差が求められない[14]．ここで異分野の用語を混在させ，LDA を使った教師なしクラスタリングが分野の区別を近似するかという課題に設定を変えた．現状でこの目的が達成されているとは言えないが，語の内部構造だけで分野が決らない場合がある事を考慮すると，それなりに妥当な結果が得られていると結論できる．

4.2　今後の展開

LDA の利点は教師なし分類が実現できるのみならず，訓練データにない事例を分類できる (何らかのエンコードを与える)．これは人手分類にはない利点である．訓練データ外の事例の分類性能の評価を将来的に実施の予定である．

4.3　結論

以上の事から明らかであるように，本研究は用語の自動分類を実現するものではない．だが，次の 3 つの重要な結果が得られていると結論して良いだろう．1) 構成要素の不連続な *n*-gram を term に使った LDA は，全体 (= document) の統語構造に相当する情報をエンコードする．2) LDA を使った用語の分類は，マクロな分類 (topic 数が小さい場合) とミクロな分類 (topic 数が大きい場合) のいずれをも実現する．

更に言えば，3) LDA は任意の大きな単位の構成関係の教師なし解析に使える．この観点で言えば，document と term という用語は LDA という解析の本質を表わしていない．document は任意の全体，term はそれらの部分であると理解するのが本質的であり，topic も (当初の想定とは違って) 意味的 (semantic) なものであるとは言えない．LDA の分析は意味的と言うより位相的 topological と考えた方が適切であるように思われる．

13)　この設定は LDA を使った各国語の文字綴りと発音の解析 [8] の異分野応用である．

14)　実は「分類に唯一の正解はない」という事 [9] は強調しておいて良いかも知れない．

謝辞

本研究は JSPS 科研費 JP21H03777 の助成を受けたものである．

LDA tuning に R Core Team の開発した R (https://www.R-project.org/) の version 4.3.x を使用した．その他のデータ解析と可視化には Anaconda 3 (https://www.anaconda.com) の 23.11.00 で Jupyter Notebook 6.5.4，Python 3.10 を実行して実現した．

参考文献

[1] 黒田航, 相良かおる. 医療用語の is-a オントロジー構築の FCA を使った効率化. 言語処理学会第 28 回年次大会発表論文集, pp. 705–709, 2022.

[2] 東条佳奈, 黒田航, 相良かおる, 高崎智子, 西嶋佑太郎, 麻子軒, 山崎誠. 実践医療用語_語構成要素語彙試案表 ver.2.0 の構築. 言語資源ワークショップ, 2022.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. **Journal of Machine Learning Research**, Vol. 3, pp. 993–1022, 2003.

[4] 岩田具治. トピックモデル. 講談社, 2015.

[5] 黒田航, 相良かおる, 東条佳奈, 麻子軒, 西嶋佑太郎, 山崎誠. 要素の重複と不連続性を扱える抽出型の語構成要素解析: 並列分散型形態素解析の提案. 言語処理学会 29 回年次大会発表論文集, 2023.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. **Journal of Machine Learning Research**, Vol. 9, pp. 2579–2605, 2008.

[7] D. Gutherie, B. Allison, W. Liu, L. Gutherie, and Y. Wilks. A closer look at skip-gram modeling. In **Proceedings of the 5th International Conference on Language Resources and Evaluation**, pp. 1–4, 2006.

[8] Kow Kuroda. Finding structure in spelling and pronunciation using Latent Dirichlet Allocation. In **Proceedings of the 30th Annual Meeting of the Natural Processing Association**, 2024. [submitted].

[9] キサク・ヨーンキャロル. 自然を名づける: なぜ生物分類では直感と科学が衝突するのか. NTT 出版, 2013. [Original: Carol Kaesuk Yoon (2009). Naming Nature: The Clash Between Instinct and Science. W. W. Norton & Company.].

Appendix

階層クラスタリング

D のランダムな部分サンプル 200 事例の階層クラスタリングの結果を示す．図 7 は topic 数が 10 で term が 1-gram の場合，図 8 は topic 数が 10 で term を skippy 2-gram にした場合である．末端の色は t-SNE の図の分野の色と同一である．

図 7　1gram を用いた LDA (topic 数: 10) の部分サンプル 200 の階層クラスタリングの結果

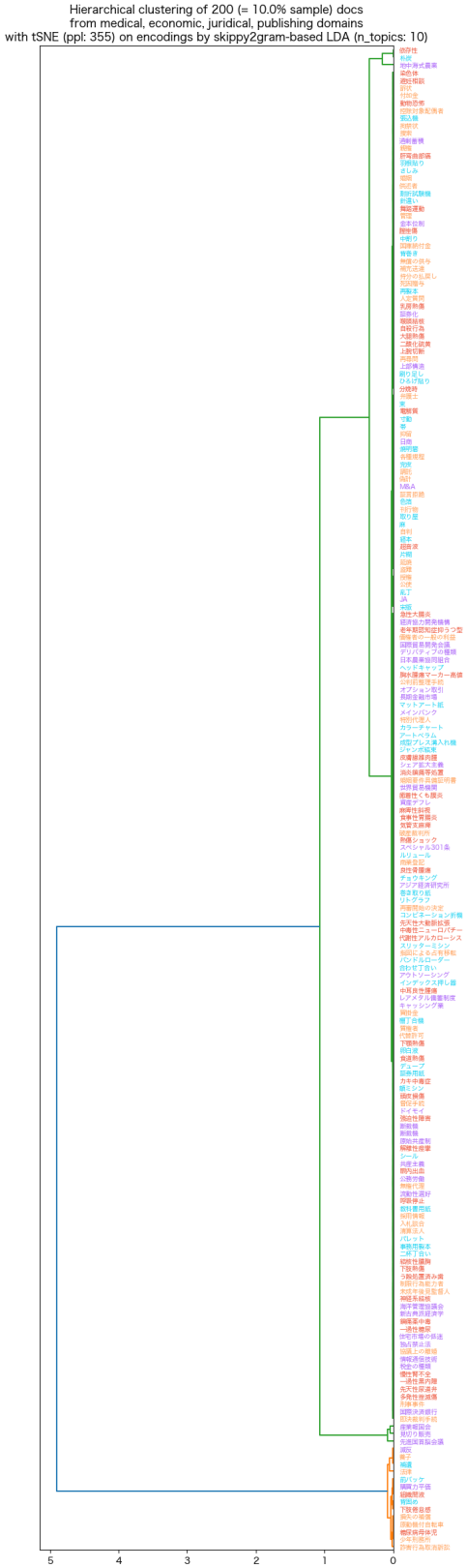

図 8　skippy 2gram を用いた LDA (topic 数: 10) の部分サンプル 200 の階層クラスタリングの結果