動的トピックモデルを用いた 特許技術専門用語に対する技術進展分析

岩田 真奈[1] 内海 祥雅[2] 松田 義郎[2] 齋藤 歩美[2] 田中 義敏[1] 中田 和秀[1]
東京工業大学工学院経営工学系[1] 楽天株式会社[2]
iwata.m.ac@m.titech.ac.jp,
{yoshimasa.utsumi,yoshiro.matsuda,ayumi.f.saito}@rakuten.com,
{tanaka.y.al,nakata.k.ac}@m.titech.ac.jp

1 はじめに

特許には，世界対応の国際特許分類 (International Patent Classification，IPC) が存在する．また，日本では，日本での特許出願が多い分野への対応などを目的として作成された，IPC をベースとして独自に細展開した特許分類である File Index(FI) も使用される．FI は約 19 万種類にも及ぶ詳細な分類である [1]．FI は，技術の進展に対応し適切なサーチキーとして機能する必要がある．そのため，特許庁により年に 1 回から 2 回，必要な分野において改正が行われる [2]. FI の改正に関しては 2006 年以降で，約 3 万件の新設と約 3 万件の廃止が行われており，大量の FI の中から改正を行うべきであるものを判断することが困難であることは容易に想像できる．そこで，改正するべき技術の候補を抽出することで，特許庁の業務補助ができると考えられる．

そのため，本研究の目的は，特許文書を基に，技術進展分析を行うモデルを作成することにより，特許庁の FI 改正という業務の補助を果たすことである．本研究では，まず，特許文書から特許技術専門用語を抽出するという作業を行う．その後に，抽出した特許技術専門用語を動的トピックモデルに適用する．最後に，特許技術専門用語の時系列変化及び実際の FI 改正結果との比較を行うことによって，動的トピックモデルの有用性を確認する．また，トピックを利用した共起単語の確認により，技術進展分析を行い，FI 改正業務の補助という観点での実用の可能性を示す．

2 関連研究

2.1 専門用語抽出

中川らは，専門用語の多くは複合語であることを利用し，専門用語を抽出している [3]．そして，専門用語にはターム性 (ある言語的単位を持つ分野固有の概念への関連性の強さ) があるという特徴に注目し，ターム性を反映した専門用語抽出法を採用した．その上で，抽出された用語候補集合における構造の情報に加え，コーパスにおける個別用語候補の統計的性質 (純粋な出現回数など) に関しても考慮したスコアを定義した．

特許の専門用語抽出 [4]，[5] に関する論文も基本的にはこの考え方を採用している．しかし，中川ら [3] の研究は特許分野に限られていないため，特許の専門用語抽出においては，特許特有の用語 (前記，当該などの特許固有の接頭辞) は除くべきだと示されている．

2.2 トピックモデル

2.2.1 トピックモデル

トピックモデルは，1 つの文書がトピック分布を持ち，1 つのトピックが単語分布を持つ仮定した文書生成モデルである．トピックモデルを用いることで，人間が文書を読むことなく大量の文書集合から話題になっているトピックを抽出することができる．トピックモデルの中でも，Blei らによって提案された Latent Dirichlet Allocation(潜在ディリクレ配分モデル，LDA)[6] が広く使われている．しかし，LDA においては時系列性は考慮されいないため，トピックの推移を確認することはできない．

2.2.2 動的トピックモデル

2.2.1 節の問題に対応しようと Blei らによって提案された Dynamic Topic Model (動的トピックモデル)[7] は LDA を拡張したモデルであり，時系列性を考慮する．動的トピックモデルにおいては，データセットは指定された時間 (タイムスライス) ごとに分割され，文書のトピック分布のパラメータ，および各トピックの単語の分布は時間とともに変化する．

しかしながら，[7] では，変分推論が行われており，モデルを更新するには全てのデータセットを読み込む必要があった．そのため，データの拡大がしにくく，大規模な問題が解けなかった．そこで，Bhadury らは，確率的勾配ランジュバン動力学法や，ギブスサンプリングの使用を提案した [8]．そして，並列可能な推論アルゴリズムを作り，高速化に成功した．その結果，大規模な動的トピックモデルを学習することができた．また，変分推論においては平均場近似という仮定が行われていた．それに対し，[8] では，このような不当な仮定を行わないことにより，低いパープレキシティも達成するなど，精度の保証もされている．

3 提案手法

本研究は，FI の改正という観点を基づき，動的トピックモデルを用いた特許公報の分析を行った．以下では，特許技術専門用語抽出，動的トピックモデルの適用，評価という流れで説明を行う．

3.1 特許技術専門用語抽出

先行研究での結果を踏まえ，特許固有の接頭辞や指示代名詞 (前記，当該など) を形態素解析前に除外するべき単語として定義した．動的トピックモデルにおいては，語彙数が増えると計算時間が増大する．また，トピックの中身の把握及び「改正すべきである技術を提示する」という目標においては，すべての単語を入力として使う必要がない．このような点を考慮し，頻出単語上位 K 単語を抽出し，動的トピックモデルの入力とすることが適切であると考えた．一方，出現頻度が多いだけの単語は情報を持っておらず，トピックモデルを適用した際にノイズとなる．そのため，頻出単語を抽出すること，及び，頻出単語で特許技術専門用語でないものを除外することが必要となる．そこで，[3] のアイデアを基に，頻出一般用語スコアを定義した．まず，複合名詞のターム性を反映したスコアを定義するために，分野における基礎概念度合い，すなわち特許技術専門用語度合いを最も反映していると考えられる連接種類数を使用した．[3] では，複合名詞のターム性を反映したスコアとして相乗平均を考えていたが，大きな概念が少しでも含まれている単語は特許技術専門用語となる可能性が高いと考え，連接種類数の最大値を複合名詞のターム性を反映したスコアとした．

また，頻出単語で特許技術専門用語でないものを除外するべきという考えを反映するべく，複合名詞 CN の頻出一般用語スコア $FNS(CN)$ を次の様に定義した．

$$FNS(CN) = \frac{f(CN)}{LR(CN)} \tag{1}$$

ここで，$f(CN)$ は複合名詞 CN の単独出現回数，また $LR(CN)$ を複合名詞 CN のターム性を反映したスコアとなる．そして，$FNS(CN)$ の上位 K 個を頻出単語であるが特許技術専門用語でないものと考え，除外単語として設定することで，ノイズを除外することができた．

3.2 動的トピックモデルの適用

それぞれの特許は様々な技術 (=トピック) で構成されており，技術は特許技術専門用語の集合のようなものであると考えられる．そのため，技術の進退を考える際に，トピックという考え方が重要となる．よって，本研究では動的トピックモデルを適用した．その詳細は Iwata et al.[9] を参照されたい．ただし [9] では，「複数，１つ」などのノイズが入り結果解釈がしづらいという問題点が存在した．その問題を解決するために，前処理として特許技術専門用語抽出を行うことにした．また，計算時間を短縮するために，[7] ではなく，[8] で提案された手法を採用した．

3.3 評価

本研究ではトピックモデルの評価という観点と，実際の改正結果との比較という観点から評価を行った．実際の改正結果との比較では，特許庁の FI の改正情報 [2] を使用した．ここで，FI の改正情報に対しても 3.1 節と同じ操作を行い，技術に関する専門用語を抽出した．この結果を基に動的トピックモデルの結果との比較を行った．

4　特許公報を用いた検証

4.1　データセット

特許技術専門用語抽出

検証に用いるデータセットは，以下の 2 種類である．

- 特許技術専門用語抽出の精度を確認するためのデータ．
  2017 年 10 月 01 日から 2020 年 10 月 01 日に出願され，かつ，2020 年 10 月 01 日時点で登録されている登録公報 (全 72,456 件)．この登録公報の中の「発明の名称」及び「要約」を使用した．
- 特許庁の公開しているデータ．
  [10] に掲載されている特許文献の日英機械翻訳辞書に登録されている単語を今回の正解とみなして利用した．登録されている単語に関しては，※などの記号を置き換えた上，それらの単語に関しては形態素解析を行うなどの処理は行っていない．全部で 108,472 件の単語が登録されている．

動的トピックモデル

2006 年 01 月 01 から 2019 年 12 月 31 に出願され，かつ，2020 年 10 月 01 日時点で登録されている登録公報で，クラスが F21(照明)，G06(計算，計数)，C12(生化学) であるものである．また，クラスの特許数にばらつきがあったため，アンダーサンプリングを行った．その結果，対象となる登録公報は 64,008 件となった．また，この登録公報の中の「発明の名称」及び「要約」を使用した．

評価

特許庁の実際の改正結果は [2] から取得を行った．2006 年以降の改正情報の中で新設と廃止を改正結果とみなし，実験の評価に使用した．

4.2　特許技術専門用語抽出

本節では，3.1 節で示した特許技術専門用語抽出を行うことにより，単純な名詞抽出よりも正しく特許技術専門用語が抽出できることを確認する．また，今回は前処理として，すべてカタカナの単語は特許専門用語であると定義している．

まず，$LR(CN)$ の定義を [3] のように相乗平均を用いた場合と，3.1 節のように最大値を用いた場合の比較を行った．すると，上位単語の中でも，相乗平均を用いた場合は「センサアンプ部, マテリアルシール部, 脱水部」などの特許技術専門用語が抽出されるが，最大値を用いた場合においてはそのようなことは発生せず，より一般用語であると考えられるものが抽出できた．

また，4.1 節で述べた特許文献の日英機械翻訳辞書を用いた比較を行った．ここで，評価指標として利用する評価指標は，Precision@N である．Precision@N とは上位 N 単語のうち正解となる単語の割合であり，推薦システムにおいてよく用いられる手法である．また，本来であれば翻訳用ではない特許技術専門用語辞書で，かつ，前処理が適切になされているものを正解データとして使うことが適切であることに留意する．また，今回は複数のしきい値で確認を行った上，動的トピックモデルに適用する数として適切な上位 3,000 単語を使用した．さらに，同様に複数のしきい値で確認を行った上，除外単語は 500 単語とした．ここで，複合名詞の考慮の効果，及び除外単語の設定の効果を確認するため，除外単語を考慮せず，単名詞のみの頻出単語 3,000 件を抽出した場合と，複合名詞も含む頻出単語 3,000 件を抽出した場合を比較した．

表 1　特許技術専門用語抽出の精度

	Precision@3000
単名詞のみ (除外なし)	2.5%
単名詞のみ (除外あり)	5.2%
複合名詞を含む (除外なし)	2.4%
複合名詞を含む (除外あり)	**5.6%**

表 1 より，単名詞のみではなく複合名詞を含む形にする効果，及び除外単語を設定する効果が確認できた．

4.3　動的トピックモデル

4.2 節の結果を基に，除外単語は 500 単語とした上で，動的トピックモデルに適用する単語数は除外されていない単語の中で頻出上位 3,000 単語を使用した．また，タイムスライスは，出願年とし，2006 年から 2019 年までの各年の時系列変化を確認した．トピック数は 15，イテレーション数は 5,000 とした．

動的トピックモデルの評価

動的トピックモデル自体の評価には，パープレキシティ [11] を用いた．パープレキシティを図 1 に示す．この結果より，イテレーション数は 1000 ほどで実験結果は十分収束していることが確認できる．また，すべての年において学習が適切に進行してい

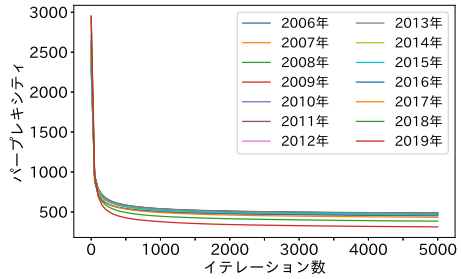

図 1　パープレキシティ

ることが確認できる．

動的トピックモデルにおいて，時間とともに変化するトピックの単語分布の例を図 2 に示す．

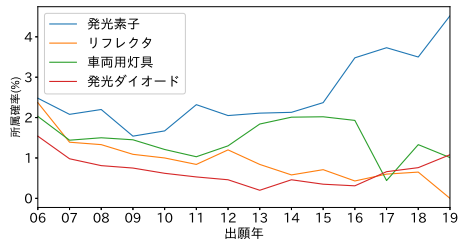

図 2　トピック 5 の中の単語の所属確率の推移

このように，各トピックにおいて構成される単語の所属確率の変化を確認することで，技術進展分析が可能となる．

4.4　単語集計の結果との比較

トピックモデルの有用性について説明するために，単語を年ごとに単純集計した結果との比較を行った．また，今回は正解データとして実際の FI 改正結果を用いたが，専門家によりアノテーションすることの方が好ましいと考えられる．単純集計のスコア $SF(CN)$ を以下の様に定義した．

$$SF(CN) = \max(\{g(CN_t)\}_{t=0}^{T}) - \min(\{g(CN_t)\}_{t=0}^{T}) \quad (2)$$

ただし，$g(CN_t)$ は，複合名詞 CN のタイムスライス t における特許 1 つに対する平均出現回数である．

動的トピックモデルのスコア $F(CN,i)$ を以下の様に定義した

$$F(CN,i) = (\max(\{p_{CN,i,t}\}_{t=0}^{T}) - \min(\{p_{CN,i,t}\}_{t=0}^{T}) \times S(i) \quad (3)$$

ただし，i はトピック ID である．$p_{CN,i,t}$ は動的トピックモデルから得られるタイムスライス t，トピック i における複合名詞 CN の所属確率である．また，$S(i)$ は，トピックの平均的な大きさであり，動的トピックモデルから計算した．動的トピックモデルの特性上，複合名詞 CN に対する $F(CN,i)$ は複数ある場合があるが，最大となる値のみ利用した．

これらのスコアを基にソートを行い，Precision@N を確認した．また，今回のデータにおいて FI の改正は多く行われているため，大きな数の N を確認した．

表 2　Precision@N

Precision@N	単純集計	動的トピックモデル
Precision@50	70.0%	**74.0%**
Precision@100	69.0%	**72.0%**
Precision@250	65.2%	**66.0%**
Precision@500	57.8%	**60.6%**

表 2 より，動的トピックモデルを使用することで，改正すべき単語を適切に把握できると分かった．

最後にトピックモデルを用いることによる今後の可能性について述べる．ここでは例として，「タッチパネル」に着目する．$F($ タッチパネル$,i)$ が最大となるトピックであるトピック 15 の上位単語に着目すると，「制御，状態，表示，検出，操作，電子機器，表示部，位置，制御部，機能，タッチパネル，表示装置」であった．それに対して，実際の改正結果においては，G06 クラスの中で「複数のタッチパネルの制御，例．連結、跨ぐ操作」というタイトルで，2014 年 4 月に改正が行われている．また，このタイトルから，提案手法の特許技術専門用語抽出作業を用いることで「タッチパネル，制御，例，連結，操作」というが抽出される．これは，トピック 15 の上位単語と重複している．このように，どのトピックに所属しているか，及びトピックの中身を確認することで技術を容易に把握でき，技術進展分析の補助となると言える．

5　おわりに

本研究は，特許文書から技術の進展分析を行い，FI の改正の補助を行うという目的で，モデルを構築した．特許技術専門用語抽出のために，複合名詞を抽出し，その後頻出一般用語を除いた．このことにより，ノイズとなる単語が減少し特許技術専門用語が正しく抽出できた．また，動的トピックモデルを適用することにより，単純な単語の推移を確認するよりも改正すべき単語を正しく発見できることが分かった．さらに，トピックモデルの特性を生かすことで，技術を特許技術専門用語の集まりとみなすことができるという有用性を確認できた．

参考文献

[1] 特許庁. 特許分類の概要とそれらを用いた先行技術調査, (2021-1 閲覧). https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_jitsumusha_txt/04t.pdf.

[2] 特許庁. Fi 改正情報, (2021-1 閲覧). https://www.jpo.go.jp/system/patent/gaiyo/bunrui/fi/f_i_kaisei.html.

[3] 中川裕志, 湯本紘彰, 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[4] 柚木山駿, 太田貴久, 小林暁雄, 増山繁. 特許関連業務支援のための技術用語自動抽出の試み. 言語処理学会, 第 22 回年次大会 発表論文集, pp. 326–329, 2016.

[5] 粟飯原俊介, 内山清子, 石崎俊. 特許文における分野オントロジー構築のための重要複合語の抽出と重要複合語間関係の定義. 言語処理学会年次大会発表論文集, Vol. 13, pp. 871–874, 2017.

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

[7] David M.and John D. Lafferty Blei. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[8] Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 381–2016, 2016.

[9] Mana Iwata, Yoshiro Matsuda, Yoshimasa Utsumi, Yoshitoshi Tanaka, and Kazuhide Nakata. Technical progress analysis using a dynamic topic model for technical terms to revise patent classification codes, 2020. https://arxiv.org/abs/2012.10120.

[10] 特許庁. 特許文献機械翻訳の辞書等の作成, (2021-1 閲覧). https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyo_dictionary.html.

[11] 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.

A　付録

4.2 節の比較において使用された上位 20 単語.

表 3　先行研究との比較

定義	上位 20 単語
先行研究	要約, 一対, こと, 課題, 場合, 否, よう, ため, 図, 発明, 複数, 提供, とき, これ, センサアンプ部, マテリアルシール部, 脱水部, 内側吸着部, 内部透過部, 先端ワイヤ部
提案	要約, 課題, 一対, こと, 場合, 具備, 発明, これ, 図, もの, 否, 提供, 際, ここ, それ, 適切, 新た, １つ, 選択図, もと

4.3 節の動的トピックモデルに適用する単語における頻出単語は以下の通り.

表 4　3 クラスの頻出単語の比較

定義	上位 20 単語
除外なし	こと, 要約, 選択図, 課題, 解決手段, 提供, 図, ため, 複数, 発明, 方法, よう, 光, 形成, 配置, 光源, 場合, 照明装置, 構成, 使用
除外あり	照明装置, 検出, 可能, 位置, 表示, 製造方法, 画像, 間, プログラム, 情報, ユーザ, 工程, データ, 細胞, 導光板, 装置, 領域, 方向, 制御, 判定