２Ａ０６

機械学習を用いた科学技術イノベーション政策における論点の抽出：一線級の研究者・有識者を対象とした大規模意識調査の自由記述を用いたチャレンジ

伊神正貫，村上昭義（文科省・ＮＩＳＴＥＰ）

1. はじめに

科学技術・学術政策研究所(NISTEP)では、第 5 期科学技術基本計画期間中の我が国における科学技術やイノベーションの状況変化を把握するため、一線級の研究者や有識者約 2,800 名を対象とした 5 年間の継続的な意識調査(第 3 期 NISTEP 定点調査[1])を 2016 年度より新たに開始した[1, 2]。

NISTEP 定点調査では、基本計画を踏まえて作成した質問に対する回答者の認識を主に問うているが、「大学・公的研究機関における研究人材の状況」等について回答者の見解を問う自由記述質問もある(図表 1 に自由記述質問一覧を例示した)。

2016 年度に実施した NISTEP 定点調査(以降、NISTEP 定点調査 2016 と呼ぶ)では、自由記述質問に、約 4,400 件(文字数約 55 万字)の回答が寄せられた。これらは、科学技術イノベーション政策における課題等の論点を抽出するには重要な情報源[2]といえる。過去の調査においては、NISTEP 定点調査の担当者が、自由記述を読み込み、目視で論点の抽出、自由記述の分類を行っていた。しかしながら、自由記述の量は膨大であり、論点の抽出・分類には多大な労力を要する。加えて、論点の抽出・分類に際して、担当者の主観を排除することは困難である。

近年、自然言語処理及び自然言語処理への機械学習の適用が目覚ましい進展を遂げており、高度な分析が可能なソフトウェアもオープンソースとして利用可能となっている。そこで、本報告では、NISTEP 定点調査から得られた自由記述から、機械学習を用いて科学技術イノベーション政策における論点の抽出を試みた結果について報告する。

機械学習を用いた自然言語処理には多様なアプローチがあるが、本報告では 1)TF-IDF 法と t-SNE による可視化、2)Word2vec と t-SNE による可視化という 2 つの方法による分析結果について述べる[3]。なお、本報告に示す結果は、初期的な分析結果であり、最終的な結果は変わる可能性がある。

図表 1 自由記述質問一覧(NISTEP 定点調査 2016 の例)

自由記述質問
大学・公的研究機関における研究人材の状況
研究環境及び研究資金の状況
学術研究・基礎研究と研究費マネジメントの状況
産学官連携とイノベーション政策の状況
大学改革と機能強化の状況
社会との関係深化と推進機能の強化の状況

2. 分析に用いたデータとその前処理

本報告では、2011, 2013～2016 年度の NISTEP 定点調査で得られた自由記述を分析に用いた[4]。また、分析を行う前につぎに述べる前処理を行った。

まず、自由記述の各回答(100 文字以上のもの 8,771 件)を、Janome[5]を用いて分かち書きした。その際、名詞及び接頭詞のみを抽出し、名詞や接頭詞が連続した場合は一続きの単語として抽出した(例: 若手研究者など)。また、一部の科学技術用語については、表記ゆれの吸収(ポスドク→ポストドクター)を行った。

つぎに、上記で得られた、分かち書きの結果を用い

1 これまでに第 3 期科学技術基本計画期間中(2006～2010 年度)、第 4 期科学技術基本計画期間中(2011～2015 年度)の 2 期 10 年間にわたって NISTEP 定点調査を実施している。第 5 期科学技術基本計画期間中(2016～2020 年度)の NISTEP 定点調査は第 3 期目となる。2016 年度調査は 2016 年 10 月～2017 年 1 月に実施し、回答率は 93.6%(回答者数 2,592 名/送付者数 2,770 名)であった。

2 NISTEP 定点調査の自由記述は、大学や公的研究機関等における現場の声であることから、文部科学省の各課室に関連する自由記述の抜粋を提供している。

3 本報告では TF-IDF 法及び t-SNE については scikit-learn(http://scikit-learn.org/stable/)、Word2vec については gensim(https://radimrehurek.com/gensim/)を用いた。

4 2012 年度調査では自由記述の質問を行っていない。

5 Janome(http://mocobeta.github.io/janome/)は Python で記述された、辞書内包の形態素解析器である。

て、質問単位の bag-of-words を作成し、質問ごとに各単語の TF-IDF 値を求めた。以降の分析では、TF-IDF 値が 0.005 以上の単語を分析に用いている。ここで質問単位の bag-of-words を用いたのは、各質問によって特徴的に使用される単語（特徴語）が異なることが経験的に分かっていたことと、質問によって回答数にばらつきがあることから特徴語を各質問から均一に取り出すためである。

以降では、これらの前処理の結果得られたデータを「分析用データセット」と呼ぶ。

3. 機械学習を用いた自由記述の可視化

ここでは、TF-IDF 法と t-SNE による可視化(3-1)、Word2vec と t-SNE による可視化(3-2)の概要について示す。なお、Word2vec や t-SNE で設定する各種パラメータについては、可視化結果を目視確認することで経験的に決定している。

3-1. TF-IDF 法と t-SNE による可視化

TF-IDF 法と t-SNE による可視化では、つぎに述べる手順で自由記述の可視化を行った。

まず、分析用データセットから NISTEP 定点調査 2016 の自由記述(2,142 件の回答)を抽出し、回答単位で bag-of-words を作成し、回答ごとに各単語の TF-IDF 値を求めた。2,142 件の回答には、1,362 のユニークな単語が含まれる。

つぎに、得られた TF-IDF 値から、各回答について文書ベクトル（以降、TF-IDF ベクトルと呼ぶ）を求めた。TF-IDF を用いた 2,142 件の回答のベクトル化から 2,142(回答数)×1,362(ユニークな単語数)のマトリクスが得られた。

本報告では、上記のマトリクスを、特異値分解を用いて 2,142(回答数)×20 に次元圧縮した後に t-SNE による可視化を行った。t-SNE は高次元のデータの可視化に適した次元圧縮テクニックである[6]。

3-2. Word2vec と t-SNE による可視化

Word2vec と t-SNE による可視化では、つぎに述べる手順で自由記述の可視化を行った。Word2vec とは 2 層のニューラルネットワークを用いて、単語をベクトル表現する手法である。2013 年に Google 社の Mikolov 等によって提案された後[3]、応用が進んでいる。

まず、分析用データセットに含まれる全ての自由記述(8,771 件、1,362 のユニークな単語)を用いて、Word2vec により単語ベクトル(300 次元)を得た。図表 2 は「女性研究者」、「若手研究者」、「研究施設・設備」、「電子ジャーナル」という単語について、Word2vec から得られた類似の単語を示した結果である。例えば、「女性研究者」の場合、「女性教員」、「女子学生」、「男女」、「男性」、「出産・育児」が類似する単語として抽出されている。他の単語についても、これらの単語と経験的に近い(共起する)と思われる単語が抽出されていることが分かる。

図表 2 Word2vec から得られた類似の単語

検索した単語	類似単語(上位5)とコサイン類似度
女性研究者	女性教員(0.59)，女子学生(0.57)，男女(0.52)，男性(0.51)，出産・育児(0.5)
若手研究者	任期制(0.53)，ポストドクター(0.49)，若手人材(0.49)，任期(0.47)，身分(0.46)
研究施設・設備	研究機器(0.52)，概算要求(0.52)，老朽化(0.5)，機器(0.48)，装置(0.46)
電子ジャーナル	高騰(0.69)，図書館(0.54)，老朽化(0.54)，閲覧(0.54)，大学図書館(0.53)

つぎに、式(1)に示したように TF-IDF ベクトル(3-1 で得たもの)で単語ベクトルを重みづけしたものの和を求めることで、各回答のベクトル表現を得た(以降、回答ベクトルと呼ぶ)。

$$\vec{v}(d) = \sum_{x \in d} w(x) \times \vec{v}(x) \qquad (1)$$

ここで、$\vec{v}(x)$は文書dに含まれる単語x の単語ベクトル、$w(x)$は TF-IDF ベクトルから得られた文書dに含まれる単語x の重み、$\vec{v}(d)$は回答ベクトルである。回答ベクトルについては、大きさが 1 となるように正規化を行った。2016 年度の NISTEP 定点調査の自由記述 2,142 件のベクトル化から 2,142(回答数)×300 のマトリクスが得られた。

最後に、上記のマトリクスを、特異値分解を用いて 2,142(回答数)×30 に次元圧縮した後に t-SNE による可視化を行った。

6 t-SNE の詳細については、つぎの URL を参照のこと。
https://lvdmaaten.github.io/tsne/ (2017 年 9 月 18 日アクセス)

4. 自由記述の可視化結果

ここでは、TF-IDF 法と t-SNE による可視化結果(4-1)、Word2vec と t-SNE による可視化結果(4-2)を示す。

4-1. TF-IDF 法と t-SNE による可視化結果

図表 3(a)に TF-IDF 法と t-SNE による可視化を示す。図表中の各点が、1 つの自由記述に対応している。可視化の結果を見ると、自由記述が集まっている部分が見られる。自由記述が集まっている部分が特定の「論点」に対応していると仮定し、そこに含まれる出現頻度が上位 10 の単語を図表 3(b)に示した[7]。

論点ごとに上位の単語をみると、1 は「博士課程後期の学生の状況」、2 は「産学官連携の状況」、3 は「大学における運営費交付金・基盤的経費や外部資金の状況」、4 は「研究施設・設備の状況」、5 は「研究資金の配分や評価・間接経費等の状況」、6 は「大学改革の状況」、7 は「基礎研究とイノベーションの状況」、8 は「科学技術イノベーションと社会の状況」に対応していることが分かる。他方で、9 については、1 位の単語でも出現頻度が小さく、上位 10 の単語から何かしらの概念を導き出すことは困難であった。

4-2. Word2vec と t-SNE による可視化

図表 4(a)に Word2vec と t-SNE による可視化を示す。TF-IDF 法と t-SNE による可視化(図表 4(b))と比較すると、自由記述が集まっている部分が明確には認識できないことが分かる。

「基盤的経費」を含む自由記述の位置を図表 4 中にクロス記号で示した。図表 4(a)と(b)を比較すると、後者では「基盤的経費」を含む自由記述が 1 カ所に集まっている一方で、前者では「基盤的経費」を含む自由記述が相対的に広がって分布している。これは、Word2vec では「基盤的経費」に類似する単語を含む自由記述についても類似する自由記述として判断されることの影響と考えられる。

5. 結果の評価と今後に向けて

本報告では 1)TF-IDF 法と t-SNE による可視化、2)Word2vec と t-SNE による可視化という 2 つの方法による NISTEP 定点調査の自由記述の可視化を試みた。

TF-IDF 法と t-SNE による可視化により、自由記述の大まかな分類は可能であることが確認された。しかし、図表 3(b)で示した論点は、図表 1 に示した NISTEP 定点調査 2016 の自由記述質問の構造とほぼ対応しており、この粒度の情報であれば、質問ごとに TF-IDF 値が高い単語を抽出することで得られる。

これまでの目視確認の経験から、一つの自由記述質問には、複数の論点が含まれていることが分かっている。したがって、上記で示した個別の論点の微細構造まで分析可能な方法論の確立が必要である。本報告では、目視によって「論点」の抽出を行ったが、図表 3(a)で得られた結果を k 平均法等でクラスタリングすることで、論点の微細構造まで明らかにできる可能性がある。また、本報告では、NISTEP 定点調査 2016 で得られた自由記述の全てを分析対象としたが、質問ごとに TF-IDF 法と t-SNE による可視化を行うことも、各質問における論点の詳細分析には有効かも知れない。

本報告の範囲では、Word2vec と t-SNE による可視化については、TF-IDF 法と t-SNE による可視化と比べて、明確な構造の把握が困難であった。本報告では Word2vec の学習に NISTEP 定点調査の自由記述を用いた。学習結果を見ると、注目する単語と共起すると思われる単語が幅広く抽出されている。これについては、明確な検索語が分からない状況での自由記述検索などに応用できると考えられる。他方で、この結果として、自由記述間の類似性が高めに判定されることで、Word2vec と t-SNE による可視化については、明確な構造の把握が困難であった可能性がある。Word2vec の学習の際に、科学技術白書や Wikipedia などを用いて、より幅広い単語を学習させることで、論点構造の把握が改善できる可能性がある。

また、本報告では触れなかったが、機械学習を用いた自然言語処理として、トピックモデルも有効な手段であることから、トピックモデルの活用の可能性も探索を行う予定である。

参考文献

[1] 科学技術・学術政策研究所: 科学技術の状況に係る総合的意識調査(NISTEP 定点調査 2016) 報告書, 科学技術・学術政策研究所 NISTEP REPORT No. 171 (2017)

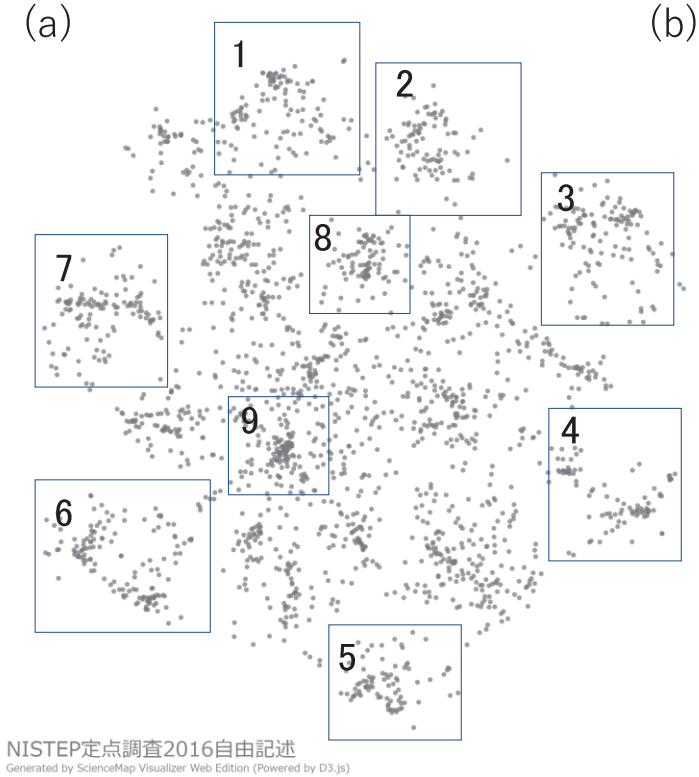
[2] 科学技術・学術政策研究所: 科学技術の状況に係る総合的意識調査(NISTEP 定点調査 2016) データ集, 科学技術・学術政策研究所 NISTEP REPORT No. 172 (2017)

[3] Mikolov, Tomas; et al.: “Efficient Estimation of Word Representations in Vector Space” arXiv:1301.3781

[7] 図表 3(b)中の「論点」については、目視によって試行的に抽出した。

図表 3 TF-IDF 法と t-SNE による可視化(a)と各部分における単語の出現頻度(b)

(a)

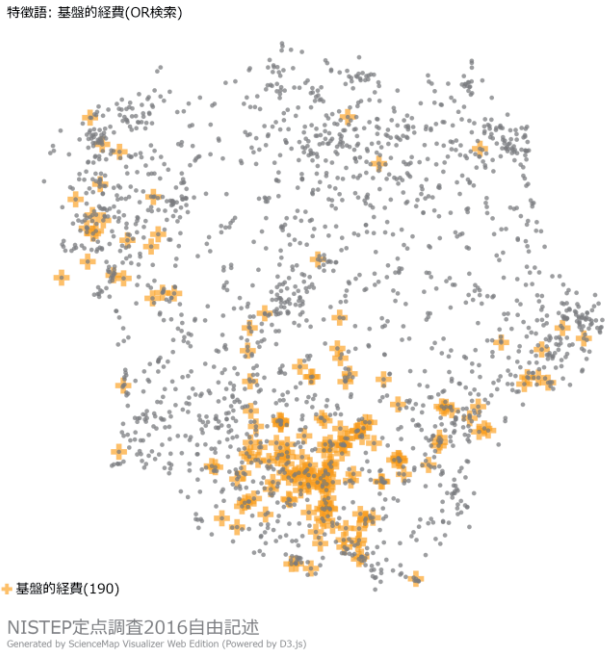


(b)

論点番号	上位の単語(出現頻度)
1	学生(167), 博士課程(94), 大学(70), 優秀(51), 進学(49), 研究開発人材(45), 研究者(33), 教育(29), 若手研究者(28), 就職(27)
2	民間企業(193), 大学(149), 連携(36), 研究者(33), 共同利用・共同研究(31), 研究開発人材(22), 知的財産(21), 産学連携(19), 実用化研究(18), 企業側(15)
3	運営費交付金・基盤的経費(159), 大学(63), 削減(49), 研究開発費(48), 競争的資金(46), 研究者(31), 外部資金(30), 予算(29), 教育(24), 時間(23)
4	研究施設・設備(86), 整備・充実(48), 研究開発費(36), 大学(35), 研究者(32), 研究環境(25), 運営費交付金・基盤的経費(17), 機器(16), 環境(15), 確保(14)
5	資源・資金配分(116), 研究開発費(96), 研究者(36), 大学(22), 基礎研究(19), 科研費(18), 評価(16), 支援(16), 予算(14), 間接経費(13)
6	改革(157), 大学(139), 学長(121), 大学改革(77), リーダーシップ(66), 機能強化(58), 執行部(56), 教員(38), 強化(26), 研究開発人材(21)
7	基礎研究(235), イノベーション(57), 研究者(37), 研究開発費(33), 実用化研究(31), 応用研究(29), 分野(28), 評価(25), 支援(25), 科研費(20)
8	社会(133), 大学(49), 科学技術イノベーション(25), 研究者(22), 理解(18), 連携(17), 関係(16), 関係深化(14), 強化(11), 教育(11)
9	連携(18), 判断(15), 分野(14), 機能(11), 政府(11), 理解(11), 失敗(10), 積極的(10), ベンチャー(9), 内容(9)

図表 4 TF-IDF 法と t-SNE による可視化(a)と Word2vec と t-SNE による可視化(b)

(a)



(b)