F-019

機械学習と可視化を用いたインタビューマイニング
Interview Mining using Machine Learning and Visualization

鄒楽晗、金川久美子、戸﨑祐輔、中藤哲也、廣川佐千男（九州大学）
Zou Yuehan, Kumiko Kanekawa, Yusuke Tozaki, Tetsuya Nakatoh, Sachio Hirokawa

1　はじめに

質的研究(Qualitative research)においてもっとも使われているデータ収集法は半構造化インタビューである。半構造化インタビューは事前に大まかな質問事項を決めておき、回答者の答えによってさらに詳細にたずねて行く簡易な質的調査法であり、構造と若干の自由度を合わせ持つことで，被面接者の語りに沿って情報を得ることが可能になるという利点がある。例えば、インタビューで得られたデータをテキストマイニングによって分析する手法は社会学や看護学において多く行われている[1,2,3]。インタビューや問答システム中の文書を横断した文間関係の分析についての研究も多く行われている。角田[4]らは、文対応推定のために文種類・文対応推定モデルを統合したモデルを提案し、従来手法よりも高い性能で文対応推定が可能であることを示した。梅村[5]らは、質問回答型電子掲示板における質問文と回答文の対応関係に着目し，提案手法を用いることによって検索精度の向上を図った.しかし、半構造化インタビューは、キーワードで質問と回答の内容を大まかに推測することは可能だが具体的な対応関係を調査する方法は未だ確立されていない。

本稿では、香港大学の Dickson Chiu らによる東アジアコレクションを保管している図書館、アーカイブと博物館の従事者に対するインタビュー事例[6]を対象として半構造化文書の分析システムを作って、インタビュー中の質問文と回答文の対応関係を分析した。具体的には、機械学習による分類と関連語の可視化を用いることにより、ライブラリアン・キュレーターの能力を表す特徴語を抽出し、必要能力を調査した。特徴語を分析システムに入力して単語関連マップを得る。これによって質問文と回答文の対応関係を分析した。

2　インタビューテキスト

2.1 東アジアコレクションおよびインタビュー形式

分析対象は、Choらの著書[6]「Inside the world’s major East Asian collections」である。本稿では、そこに掲載された第１章から第3章までのインタビュー記録を分析した。インタビューを受けたのは、東アジアコレクションを保管している図書館、アーカイブと博物館の専任職員であり、単純なアンケート形式ではなく、柔軟性と流動性を持つ半構造化インタビュー形式で行った。事前にインタビュー・ガイド(interview guides)を準備し、回答者の背景を基づいて、質問期間中、インタビュアーが興味をもった部分を追求するために新たな質問を出すことができる。これらはインタビュー調査の一貫性を維持する目的で実施された。

2.2 機械学習による「必要能力」文の特徴語の抽出

本稿は、イブラリアン・キュレーターの必要な能力を事例として分析した。特徴語が他のキーワードと共起関係を可視化するため、その最初の３つの章だけだがテキストマイニングによる「必要能力」文中の特徴語を抽出した。インタビュー記録のPDFファイルからテキストデータのみを手作業で識別し、質問文と回答文をそれぞれ文単位分けて、ライブラリアン必要の能力が表示するの文をラベルに付け加える。

機械学習[7]による質問文と解答文の識別を行った。表１に示すとおり、対象データの総文数は565個であり、総単語数は1618語である。その中に、回答文が426個であり、質問文が139個である。特に、回答文はライブラリアンの能力についての文を識別し、426個の回答文のうち57の文が抽出できた。SVMアルゴリズムによる必要能力を表示する文書群を機械的に判別し、抽出した特徴語を図１に示す。

表 1 データ概要

章	文の数	単語の数	回答文	質問文	必要能力の文
123	565	636	426	139	57

1	-0.1205	-0.1205	14	members	-2.2232 -2.2232	34 they
2	-0.2627	-0.2627	15	new	-2.1238 -2.1238	16 ccscs
3	-0.3781	-0.3781	10	lis	-1.9848 -1.9848	25 china
4	-0.4753	-0.4753	6	ability	-1.8199 -1.8199	4 groups
5	-0.4853	-0.4853	10	challenges	-1.8194 -1.8194	420 the
6	-0.5022	-0.5022	3	specialized	-1.8000 -1.8000	3 getting
7	-0.5055	-0.5055	10	since	-1.7995 -1.7995	31 use
8	-0.5134	-0.5134	119	is	-1.7513 -1.7513	33 british
9	-0.5212	-0.5212	19	jobs	-1.7390 -1.7390	120 collections
10	-0.5306	-0.5306	9	involved	-1.7390 -1.7390	21 currently

図 1 特徴語

3　質問文と回答文の対応関係

質問文と回答文の対応関係を分析するために、単語の共起関係を可視化する検索エンジンを構築した[8]。キーワードを入力すると、単語の共起関係の色を付けたマップを得ることができる。まず頻度の低い単語から始め、それよりも頻度の高い単語で、最も関連の強い単語一つ選び枝を延ばすことを繰り返し、ツリーを構成している。一つの「文」或いは一つの質問とそれへの回答の文の全部まとめた文書の「Q&A の組」、この二つの検索対象がある。

以下からは、質問文或いは回答文中の単語から自身を言及しているの内容及び質問文と回答文の対応関係の分析事例に分けて考察する。

3.1　質問文グラフ

まず、「Q&A の組」を検索対象として、質問中の単語の共起関係を表示したマップによって、それを導きだした質問文を考察する。ライブラリアンの技能経験を表すキーワード「skills」入力すると、質問のキーワードが 20 個を検索して、図２に「skills」の関連語順にマップができる。skills グラフは skills から knowledge、jobs と needed が共起している。グラフによると、質問者はライブラリアンとして必要な仕事の経験、言語知識と獲得した学位について質問すると推定できる。これは実際の質問文と一致している。具体的には、skills をキーワードとするの質問文は「Could you please describe a list of technical skills (including language skills) and professional knowledge that you require for your job as a librarian at the CCSCS? 」とあり、単語の共起関係と符合している。

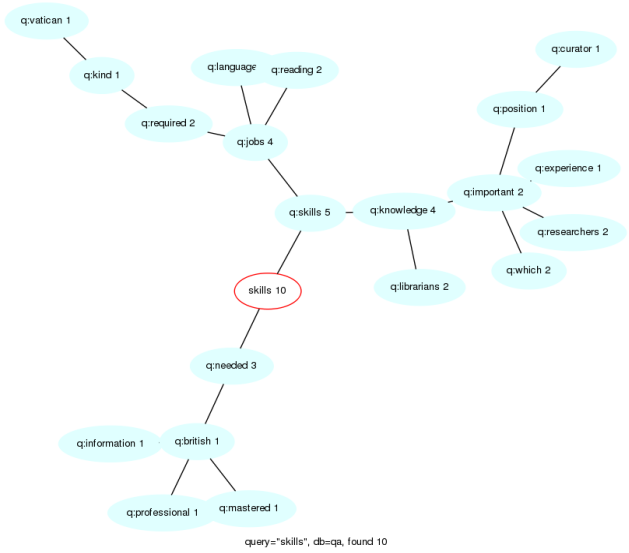

図 2 質問文グラフ

3.2 回答文グラフ

次は、「Q&A の組」を検索対象として、回答中の単語の共起関係を表示したマップによる、それを導きだした回答文を考察する。ライブラリアンの技能経験を表すキーワード「skills」入力すると、回答のキーワードが 20 個として検索する、図３のような「skills」の関連語順を現したマップができる。このマップの中に、language から interpreted、cataloguing と materials が共起している。翻訳、カタログと資料の整理とするとき、言語知識は欠かせないと推定できる。具体的には、skills をキーワードとするの回答文は「Chinese language is vital for my daily work, especially when cataloguing and selecting materials for the new acquisitions, but also when communicating with partners in China. 」とあり、単語の共起関係と符合している。

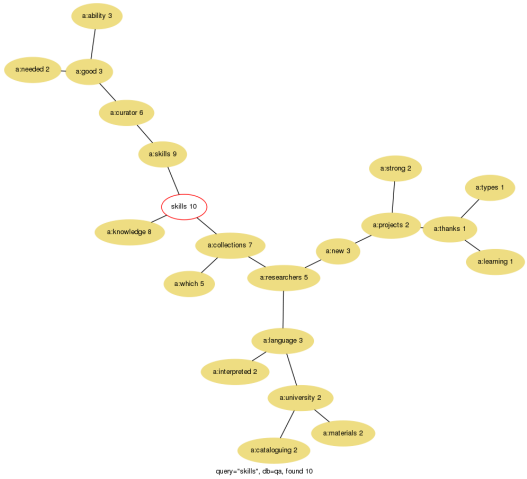

図 3 回答文グラフ

3.3 質問文と回答文の対応関係

単色の関連マップによる、質問文と回答文中の単語から自身を言及しているの内容を判断することができるとはいえ、図 4 に質問文と回答文の対応関係を探しみたい。「文」の検索対象として、「skills」入力すると、質問と回答中の単語の共起関係を表示してのマップができる。マップの両側に共通する単語は skills、language、knowledge、required と researchers がある。これによって、この対話の話題は研究者の必要能力としての言語と知識を推定できる。質問単語中の「listed」と「including」によって、キュレーターの必要能力を挙げていると推測できる。回答単語中の「important」によって、この対話は必要能力の重要性に関する問題を討論すると推測できる。推測の結果は実際のインタビュー内容と一致しています。

具体的には、質問文は「Language Skills, (2) Knowledge in History, (3) Research Skills or (4) Knowledge in Archaeology. According to your experience, which has proven to be the most important in order to do your job well? 」、回答文は「 All the skills you mentioned are very important for our work since we are dealing with collections which can, and need to, be interpreted from many different aspects. We can look at an item from a language, historical or archaeological perspective, so I would say that knowing language, history and archaeology and having research skills are all important abilities for being a good curator. 」である。

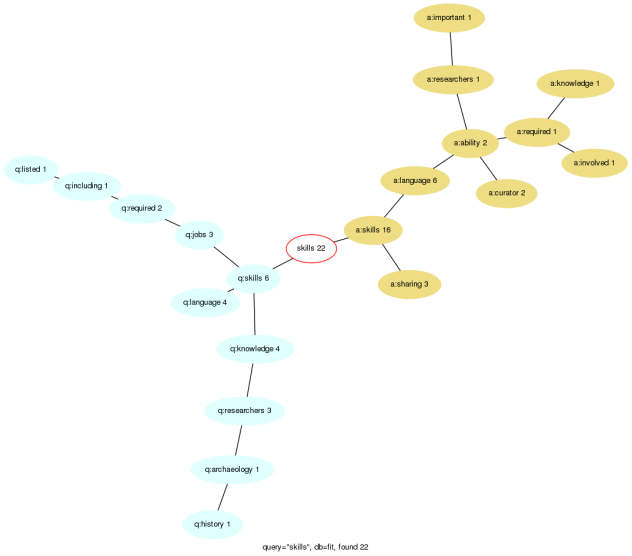

図　4 質問文と回答文の対応関係

4　質問中の単語と回答中の単語の関係

質問と回答に同じ単語が出て、同じ意味を表すの場合は多い。しかしながら、同じ単語でも q & a 文脈で異なる意味を持つ場合もある。「Q&A の組」を検索対象として、「collections」入力すると、質問と回答のキーワードはそれぞれ 5 つを検索して、質問と回答中の単語の共起関係を表示してのマップができる。図 5 に collections を含む 23 個の質問、36 個の回答がある。図中に右側に質問を表示する部分は、「east」「British」と「institution」を含む。Collections は博物館や美術館等の所蔵作品群のような意味を表すと推定できる。図中に左側に回答を表示する部分は、「materials」や「cataloguing」を含む。Collections は具体的な本や資料のような意味を表すと推定できる。具体的には、質問文は「Could you give examples of typical reference/research enquiries issued by the researchers or scholars at the East Asian Collections at the British Library? 」、回答文は「Research enquiries are mostly about our rare books collection, given the variety of the catalogues involved. Sometimes researchers based outside the UK need to preorder an item for which the shelf mark is needed. Being located outside of London they don’t have access to the cards and microfiches catalogues, so we help them to locate the material they need. 」である。推測の結果は実際のインタビュー内容と一致している。

5　まとめと今後の課題

今回は半構造化文書の分析システムを作って、インタビュー記録中のライブラリアンの必要能力の事例に基づいて、SVM によって特徴語を抽出し、共起関係を表したマップを利用して半構造化インタビューの文と単語の関係を分析する。また、これらの分析手法を他のインタビュー資料を分析対象として、ライブラリアンの能力だけでなく、一般的な質問パターンの具体的な対応関係を分析する予定である。

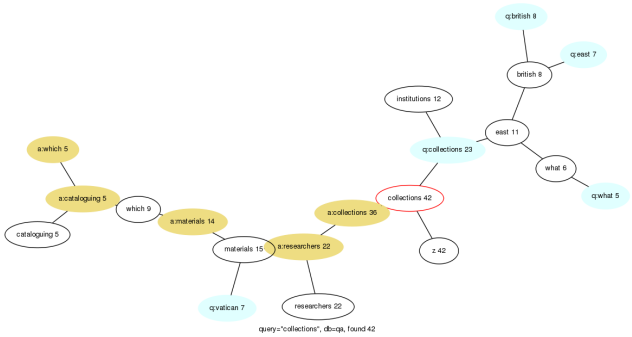

図　5 質問中と回答中の単語の区別

参考文献

[1] 加藤千佳、城丸瑞恵、いとうたけひこ, “テキストマイニングを用いた病棟看護師の実習指導に対する語りの分析,” 昭和大学保健医療学雑誌, no. 8, pp. 23-33, 2011.

[2] 渡辺裕一, “行政職員から a 現状と課題：行政職員へのインタビューに対するテキストマイニング分析から,” 武蔵野大学人間科学研究所年報 = Annu. Bull. Musashino Univ. Inst. Hum. Sci., no. 4, pp. 69-80, 2015.

[3] S. L. Hunt and C. J. Bakker, “A qualitative analysis of the information science needs of public health researchers in an academic setting,” vol. 106, no. April, 2018.

[4] 角田孝昭、乾孝司、山本幹雄, “対をなす二文書間における文対応関係の推定” Journal of Natural Language Processing 22(1), 27-58, 2015

[5] 梅村和宏、鈴木優、川越恭二, “質問文と回答文の対応関係を考慮した質問回答型電子掲示板検索手法” 情報処理学会研究報告データベースシステム (DBS) 2006(78(2006-DBS-140)), 343-350, 2006-07-14

[6] A. Cho, P. Lo, and D. K. W. Chiu, Inside the world’s major East Asian collections：one belt, one road, and beyond. Chandos Pub., 2017.

[7] Y. Adachi, N. Onimura, T. Yamashita, and S. Hirokawa, “Standard Measure and SVM Measure for Feature Selection and Their Performance Effect for Text Classification,” in Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, 2016, pp. 262-266.

[8] S. Hirokawa, B. Flanagan, T. Suzuki, C. Yin, Learning Winespeak from Mind Map of Wine Blogs, Proc. HIMI 2014, Part II, LNCS 8522, pp. 383-393, 2014