ARG WI2 No.8, 2016
WI2-2016-14

機関リポジトリから収集した学術論文のテキスト解析に関する一検討

岡本 一志
電気通信大学 大学院情報理工学研究科
kazushi@uec.ac.jp

概要 日本国内の機関リポジトリから約 45 万件の学術資料を収集し，それらに対するテキスト解析に取り組む．収集した学術資料から抽出した名詞語について，語長分布の調査や word2vec を用いた類似語の探索実験を行い，機関リポジトリで配信されている学術資料の知識発見や仮説生成のための情報源としての可能性を検討する．

キーワード 機関リポジトリ，学術資料，テキストマイニング，単語の分散表現

1 はじめに

学術論文の役割は，研究者の学術研究の成果を社会に発信することにある．近年では，その内容を必要とする人への情報提供だけでなく，計算機を用いた言語解析による知識発見や仮説生成のための情報源としての役割も期待されている．例えば，生命科学分野では文献データベース中の標題や抄録から特定の病気への関連が疑われる物質を発見する研究などが展開されている [1]．しかしながら，学術論文の言語解析による知識発見・仮説生成型の研究の大半が生命科学分野の研究であり，様々な分野の論文を横断的に活用する研究は行われていない．

その一方で，様々な学術分野の論文・国際会議の予稿原稿・博士論文などの学術資料が著者の所属機関のリポジトリを通じて無償公開されるようになってきており，その数は国内において 100 万件以上にも上る [2]．本研究では，国内の 69 の機関リポジトリから様々な学術分野の学術資料（約 45 万件）を収集し，それらから本文を抽出しそのテキストを解析することで，機関リポジトリにある学術資料が知識発見や仮説生成のための情報源として活用可能かを検討する．本研究で行うテキスト解析として，(1) 収集した学術資料ファイルから抽出した名詞語の語長分布；(2) 抽出した名詞語を word2vec[3] により特徴空間に埋め込み特徴空間上での近傍探索による関連する専門用語のアナロジータスク，を想定する．

2 機関リポジトリからの学術資料の収集

本研究ではインターネットを介して機関リポジトリで無償公開されている学術資料を分野を問わず収集する．機関リポジトリで公開されている学術資料ファイルの URL は OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting）と呼ばれるプロトコルを通じて取得することができる．本研究で学術資料を収集する機関リポジトリは DRF の WEB サイト [4] で OAI-PMH ベース URL を公開している 69 機関とする．

69 の機関リポジトリが公開している学術資料ファイルの収集手順は次の通りである：(1) 各機関リポジトリに対して，OAI-PMH の ListRecords メソッドを実行する．実行パラメータは，from=2004-01-01，until=2015-08-31，metadataPrefix=junii2 としている；(2) ListRecords メソッドによる各リクエストでは，学術資料本文ファイルへのフルテキスト URL が含まれた XML データが機関リポジトリから返却されるため，その URL を記録する．本研究では，PDF ファイルのみを対象としているため，PDF の拡張子が含まれる URL のみ記録の対象とする；(3) 記録した各 URL について GNU Wget や cURL などのダウンロードプログラムを実行し，PDF ファイルを保存する．収集した PDF ファイル数は 449,029 ファイルであり，総ファイルサイズは約 1TB であった．

3 収集した学術資料のテキスト解析

ダウンロードした PDF ファイルに pdftotext プログラムを適用し，ファイルから改行文字で区切られた文の集合を取得する．pdftotext のオプションは “-raw -nopgbrk” としている．取得した各文について，(1) 日本語形態素解析器 MeCab[5] の適用；(2) 各文をスペースで区切られた単語の集合に分解；(3) 名詞語以外を (2) の集合から除外，の 3 処理を行い名詞語のみを抽出する．

3.1 抽出した名詞語の語長分布の分析

抽出した名詞語の語長の分布は，1 文字：1,045,790,440 語，2 文字：554,208,228 語，3 文字：69,904,962 語，4 文字：35,536,684 語，5 文字：20,713,928 語，6 文字：15,823,518 語，7 文字：11,505,889 語，8 文字：9,311,674 語，9 文字：6,313,534 語，10 文字以上：14,338,152 語となっている（同じ名詞語は重複して数えている）．これより語長が 1 文字の名詞語が最も出現していることが確認できる．その原因として，(1) 印刷媒体をスキャンし

表 1 入力語に対してコサイン類似度が最も高い 7 語

入力語	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6	rank 7
TCP	iSCSI (0.738)	HTTP (0.731)	IP (0.72)	VoIP (0.712)	DCCP (0.708)	SCTP (0.708)	IPSec (0.699)
DCT	メルケプストラム (0.624)	ウェーブレット (0.623)	WDCT (0.619)	ケプストラム (0.613)	IDCT (0.611)	FFT (0.599)	Papoulis (0.599)
BCI	FES (0.644)	SSVEP (0.633)	MEG (0.622)	ERP (0.578)	EEG (0.574)	ECoG (0.572)	イメージングデバイス (0.564)
フコイダン	マンナン (0.718)	オキナワモズク (0.704)	ラフィノース (0.69)	フロロタンニン (0.680)	セルラーゼ (0.677)	グルコシダーゼ (0.675)	ホマリン (0.674)
塩基	アミノ酸 (0.703)	グアニン (0.648)	ヌクレオチド (0.625)	シトシン (0.625)	スフィンゴイド (0.624)	アデニン (0.62)	カルシウムスルホネート (0.619)

PDF ファイルを作成する場合に OCR が適用されるが，OCR の精度が悪く，スペースなどの区切り文字が語中に含まれた形で文字列が PDF ファイルに記録されてしまうこと；(2) PDF の仕様上，行末の語の途中で改行される場合があり，意味的には行末の語と次の行の先頭の語には連続性があるものの，形態素解析にあたってはその連続性を無視していること，が考えられる．そのため，機関リポジトリから収集した学術資料を知識発見や仮説生成のための情報源として活用するためには，抽出したテキストに対する前処理が重要と言える．一方で，機関リポジトリの学術資料を機械的に活用してもらうためには，機械可読な形式でのデータ提供が望ましい．

3.2 word2vec による専門語のアナロジータスク

word2vec[3] は学習データ（スペースで区切られた単語の集合）に基づき単語を n 次元の特徴ベクトルで表現する手法である．関連語が特徴空間上で近傍にある前提のもと，word2vec 用いることにより専門語のアナロジータスクを行うことができる．本研究では，機関リポジトリから収集した学術資料が知識発見や仮説生成のための情報源として活用可能とするならば，アナロジータスクで適切な結果を得られるという前提で検討を進める．

本研究では，ダウンロードした PDF ファイルから抽出した名詞語に word2vec を適用し，名詞語を n 次元の特徴ベクトルとして表現する．そして，n 次元の特徴ベクトルで表現された全ての名詞語についてコサイン類似度を計算する．word2vec のパラメータは，次元数 $n = 200$ 以外はプログラムのデフォルト値を用いる．

通信や化学などに関する 5 つの専門語（TCP，DCT，BCI，フコイダン，塩基）について，類似度が最も高い 7 語についてまとめたものを表 1 に示す．表 1 の括弧内の数値は入力語に対するコサイン類似度の値を意味している．表 1 より，全体的な傾向として，入力語に対して何らかの関係のある語が上位 7 位までに出現していることが読み取れ，一部の例であるものの，アナロジータスクにおいて適切な結果を得られていると考える．word2vec による名詞語の特徴空間への埋め込みにおいては，ある専門語について特徴空間上での近傍に関連語が存在することを示唆していると考える．

4 おわりに

国内の 69 の機関リポジトリから様々な学術分野の学術資料（約 45 万件）を収集し，それらから本文を抽出しテキスト解析を行っている．具体的には，(1) 収集した学術資料ファイルから抽出した名詞語の語長分布の調査；(2) 抽出した名詞語をについて word2vec を用いた類似語の探索実験，を実施している．アナロジータスクの結果より，機関リポジトリで配布されている学術資料は知識発見や仮説生成のための情報源としての可能性を秘めていると考える．また，情報源として機械的な利用の際には，どのように前処理を行うかが重要といえる．

今後は，n 次元の特徴空間に埋め込まれた専門語同士の関係の可視化や，wikipedia などの他の情報源を用いた場合との比較に取り組んでいきたい．

謝辞

本研究は JST 科学技術人材育成補助金テニュアトラック普及・定着事業ならびに電気通信大学平成 27 年度研究活性化支援システムの支援を受けて実施したものである．

参考文献

[1] Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., and Ananiadou, S.: Discovering and visualizing indirect associations between biomedical concepts, Bioinformatics, Vol. 27, No. 13, pp. i111–119, 2011.

[2] Institutional Repositories DataBase コンテンツ分析システム，http://irdb.nii.ac.jp/analysis/index.php, 2015 年 11 月アクセス．

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, arXiv, arXiv:1301.3781, 2013.

[4] デジタルリポジトリ連合（Digital Repository Federation），http://drf.lib.hokudai.ac.jp/drf/index.php, 2015 年 11 月アクセス．

[5] 工藤 拓，山本 薫，松本 裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告自然言語処理，Vol. 2004，No. 47(2004-NL-161)，pp. 89–96，2004.