学術文献のテキストマイニング
言語処理学会年次大会 19 年分の予稿データの知的資産としての活用可能性の検討

那須川 哲哉　　西山 莉紗　吉田 一星
日本アイ･ビー･エム株式会社 東京基礎研究所

1. はじめに

単体として重要な知識源である学術文献を集合体として扱うことによって得られる付加価値と、その活用可能性を検討する。

文献の電子化が進んだ結果、検索技術を適用することで、膨大な数の文献の中から、自分に興味のありそうな文献を特定することが比較的容易になってきている。検索技術が可能にするのは、目を通すべき文献の候補を特定することであり、基本的には、特定された文献に目を通して何らかの知識を得ることが最終目的となる。それに対し、大量の文献に含まれている情報を組み合わせることで、個々の文献だけからでは得られない知識の獲得を目的とした技術がテキストマイニングである。[1,2]

テキストマイニングで得られる知見は、対象となるデータと分析目的によって大きく異なり、全く同じデータから、分析者次第で多種多様な知見を導き出すことができる。

本稿では、言語処理学会年次大会19年分の予稿データを対象としたテキストマイニングを実践した概要と、その活用可能性を示す。

2. データ処理の概要

筆者らは、言語処理学会20周年特別企画のオーガナイザからの配布テキストである収録論文一覧表(meta131108_1.tsv)と論文pdfからの抽出テキスト(nlp_annual_txt_20131118.tar.gz)を分析対象とし、このデータをIBM® Content Analytics with Enterprise Search Version 3.0 (以下、ICAと略記)に読み込ませた。その際、自然言語処理を適用するテキスト部分は、下記の3種類に分けて設定し、各々の箇所に頻出する表現などを分析できるようにした。

- タイトル
  収録論文一覧表中のタイトル部分
- イントロダクション
  論文pdfからの抽出テキスト中で、「１．」や「I.」のようなセクション番号に「はじめに」「始めに」「まえがき」「概要」「要旨」が続いている部分から、次のセクション番号の直前までを機械的に特定した部分
- 本文
  論文pdfからの抽出テキスト中で、タイトルとイントロダクションを除いた全て

分析対象論文数は、4079件であり、そのうち上述のロジックでイントロダクション部分を切り出せたのは3547件であった。

収録論文一覧表のメタデータにおいて、著者情報に関しては、全共著者と所属がひとかたまりになって一つのフィールドに入っている。このフィールドから、筆頭著者名と共著者名及び各人の所属機関を切り出した。さらに、所属機関を、下記3タイプに分類し、筆頭著者所属タイプ及び共著者所属タイプとして、各論文にタグ付けした。

- 大学及び高等専門学校
- 公的研究機関
- 企業

この処理により、例えば、企業研究者が筆頭著者である論文の特徴を分析することが可能になる。また、これと同様の分析効果を狙いとして、言語処理学会のホームページの年次大会優秀賞・若手奨励賞一覧を参照し、受賞論文へのタグ付けを行った。

テキスト部分からの情報抽出に関しては、ICAの基本機能[3]を活用した。ICAでは、日本語のテキストに対し、形態素解析と構文解析を行っており、解析結果の構文木に出現するパタンを検出し、加工して情報抽出することができる。このパタン検出機能を用いて、後述の特長表現[4]の抽出なども行なった。

3. 言語処理結果の概要

言語処理による情報抽出結果の例として、タイトルを形態素解析し、自立語に関して活用を終止形にするなどの処理をかけ、単語レベルのキーワードとした出力の集計結果を表1に示す。

表1から、単語レベルのキーワードを単純に頻度順に並べただけでは、意味のある情報を得ることは難しいことが分かる。ここで示されているのは、言語処理学会が対象としている大まかなレベルでの研究トピックに関連した表現であり、このレベルの情報は約20年の間にあまり変化していないことが分かる。

より詳細な内容を把握するため、名詞と判断された形態素解析で複数連続している表現をタイトルから抽出した結果を表2に示す。このレベルになると、約20年の間に多少なりとも変化している様子が見受けられる。同様の名詞連鎖をイントロダクションの部分から抽出した結果を表3に示す。

表1: 各データ集合においてタイトルから抽出した自立語の頻度上位10語 （括弧内は該当表現を含む論文数）

全データ	1995 年のデータ	2013 年のデータ
用いる (609)	用いる (15)	用いる (41)
日本語 (391)	日本語 (14)	抽出 (25)
抽出 (383)	手法 (9)	日本語 (22)
基づく (336)	翻訳 (9)	基づく (21)
自動 (310)	辞書 (8)	構築 (20)
表現 (295)	モデル (7)	推定 (20)
情報 (292)	解析 (7)	自動 (20)
コーパス (280)	文章 (6)	コーパス (18)
分析 (256)	処理 (6)	手法 (17)
利用する (235)	名詞 (6)	生成 (16)

表2: 各データ集合においてタイトルから抽出した名詞連鎖の頻度上位10語 （括弧内は該当表現を含む論文数）

全データ	1995 年のデータ	2013 年のデータ
機械学習 (30)	日英機械翻訳 (3)	機械学習 (6)
自動抽出 (28)	電子化辞書 (3)	自動抽出 (4)
自動獲得 (26)	言語知識 (2)	統計翻訳 (3)
自動構築 (25)	クラスタリング (2)	トピックモデル (3)
自動生成 (24)	対訳テキスト (2)	自動生成 (3)
係り受け解析 (23)	多義性解消 (2)	単語難易度 (3)
クラスタリング (20)	相互情報量 (2)	machine translation (2)
質問応答システム (19)	日本語文章 (2)	現代日本語書き言葉均衡コーパス (2)
音声対話システム (18)	information retrieval (1)	顔文字 (2)
質問応答 (17)	English corpus (1)	クラスタリング (2)

表3: 各データ集合においてイントロダクションから抽出した名詞連鎖の頻度上位10語（括弧内は該当表現を含む論文数）

全データ	1995 年のデータ	2013 年のデータ
提案手法 (295)	言語処理 (7)	提案手法 (26)
評価実験 (200)	単語間 (5)	機械学習 (17)
機械学習 (185)	筆者ら (5)	言語資源 (13)
先行研究 (164)	係り受け解析 (5)	情報抽出 (12)
Web 上 (160)	翻訳結果 (5)	先行研究 (11)
言語処理 (147)	意味解析 (5)	評価実験 (11)
情報抽出 (131)	複合名詞 (4)	関連研究 (9)
検索結果 (124)	格フレーム (4)	分類器 (9)
筆者ら (119)	深層格 (4)	ウェブ上 (9)
質問応答 (113)	対応関係 (3)	インターネット上 (9)

（表1表2表3において、頻度上位10語目と同頻度の表現が他に存在するケースもあるが、スペースの都合上割愛した。）

表2と表3を比較すると、タイトルとイントロダクションでは、異なる表現が用いられていることが分かる。イントロダクションで出現頻度が高いのは、「提案手法」や「先行研究」など技術文献一般に共通した表現になるが、多少頻度が低い表現の中には、「意味解析」や「インターネット上」など、研究の目的や背景を示す表現が認められ、約20年の間の変化を見出せそうである。

4. 分析内容とその活用可能性

「はじめに」で述べたように、テキストマイニングで得られる知見は、分析目的に依存する。テキストマイニングを活用する上では、データから何を分析したいかという期待と、データから実際に何が導き出せるかという現実をすり合わせ、期待と現実のギャップを埋めることが重要である。そこで、幾つかの活用可能性を考え、それに沿った分析で実際に有用な知見を得られるかを調査した。

4.1. 技術の特徴の分析

新しい研究分野を模索したり、特定の技術の適用可能性を検討したりする場合、検討対象とする技術に関して、研究の主眼や特徴的な課題を把握することには意義があると考えられる。そこで、技術名称に相関の高い表現を分析することを試みた。

具体的には、ICA で特定の技術名称を含む論文を検索し、[検索結果の論文集合における出現比率]が[全データにおける出現比率]よりも高い表現から順に表示する機能を利用して、特定の技術名称を含む論文に目立つ（出現比率が高い）表現を確認した。

その結果、例えば、「機械翻訳」に対しては、単語レベルであれば、

- 目的言語
- 原言語
- 訳文

といった表現が浮かび上がり、名詞連鎖の場合は、

- 翻訳精度
- 翻訳規則
- フレーズテーブル

といった表現が浮かび上がった。さらに、特長表現を対象とした場合は、

- 翻訳精度…向上する[1]
- BLEU 値…向上する

といった表現が浮かび上がるようになり、着目すべき表現のタイプを工夫することで、「機械翻訳で重要なことは翻訳精度や BLEU 値を向上させることである」という知見を比較的容易に得ることができそうである。

同様にして「SVM」と相関の高い名詞連鎖は、

- 分離平面
- rest 法
- カーネル関数

であり、特長表現は、

- F 値...向上する
- 認識性能...優れる
- 傾向...学習することができる

[1] 「…」は基本的に助詞などの付属語とマッチした部分である。この情報抽出には構文解析結果を用いているため、「…」の部分に、文節が入っている表現（例えば、「翻訳精度を劇的に向上する」）からも「翻訳精度…向上する」が抽出される。

であった。「LDA」を対象にした場合、相関の高い名詞連鎖は、

- トピックモデル
- 潜在トピック
- ディリクレ配分法

であり、特長表現としては、

- 生成尤度...高める
- 高精度化...期待できる
- パープレキシティ...削減できる

などの相関が高かった。これらの高相関表現は、専門家が見れば自明なものが多いが、逆にこのことは専門外のトピックをサーベイする際の有用性を示唆している。サーベイ論文のように人が精査した情報源と比較すると、ノイズが多く、内容の解釈を必要とする情報源であるが、着目すべき表現を自動的に提示する点で有用性が認められ、提示された表現を含む論文を容易に参照できれば、対象分野に関する知識を効率良く得られると考えられる。テキストマイニングにおいては、こうして見出された表現を含む文献に分析対象を絞り込み、原文を確認したり、さらなる深掘りをしたり試行錯誤する機能が実用上重要な役割を果たす。

4.2. 組織の特徴の分析

著者名の所属機関の情報を用いることで、学生が進路を検討する時など、自分が身につけた専門性を活かしたり、興味のあるテーマに取組めそうな研究機関を捜したりするための知識源として文献データを活用できる可能性がある。

2 節で述べた所属タイプのタグを活用して具体的に試行した結果、例えば、

「機械翻訳」という表現を含む論文に関しては、A 社の社員が筆頭著者の論文における「機械翻訳」の出現率が、全体と比較して 1.5 倍程度と、企業の中では最も比率が高く、「機械翻訳」に比較的集中した取組みをしている可能性が考えられる。但し、総件数は 6 件であり、しかも 2007 年の論文が一番新しいことから、現在は取組んでいない可能性も考えられる。

といった知見や

「テキストマイニング」という表現を含む論文に関しては、B 社社員が筆頭著者になっている論文の「テキストマイニング」出現率が全体と比較して 3.6 倍と最も高く、全 9 件ではあるが、2000 年以降 2012 年まで比較的コンスタントに出ている。したがって B 社ではテキストマイニングに取組める可能性が高いと考えられる。

といった知見が得られた。

4.3. 受賞論文の特徴の分析

論文の質を上げるための参考にするという点で、受賞論文の特徴を分析することに意義があると考えられる。但し、今回対象とした 4079 件の論文中、受賞論文は 85 件に過ぎず、その内容も幅広いため、単純な分析では明確な特徴を見出すことが困難であった。

受賞論文に相関の高い表現としては、名詞連鎖では上位 6 件が

- argmaxy (5)
- ラベルなしデータ (7)
- 損失関数 (5)
- 提案法 (8)
- 開発データ (6)
- 従来法 (8)

（括弧内の数値は、その表現を含む受賞論文の件数）である。強いてあげれば、「*統計的手法を用いて実験をし、提案手法の優位性を従来手法と比較して示している論文が受賞しやすい傾向にある*」ということが言えるかもしれないが、仮説にすぎない。但し、テキストマイニングで重要なことは、こういった仮説を見出すことであり、通常、テキストマイニングで得られた仮説の検証には、テキスト以外の情報も使う必要があることから、こういった仮説につながる傾向を見出せる点で、有用性が認められる。

特長表現の中には、受賞論文に相関の高い表現として、出現比率が 2 倍を超えるものは、

- 対象...表現できる (2)
- 成果...利用できる (2)
- 情報...考慮できる (2)
- フレーム...構築できる (2)

（括弧内の数値は、その表現を含む受賞論文の件数）のみであり、件数が少ないため、ここから何かを導き出すのは困難と考えられる。

4.4. 増加傾向の内容の分析

動向を探るという観点では、どのような表現が増加傾向にあるかに関心が寄せられることが多い。ICA には最近の増加傾向が高い表現を特定する機能があり、

タイトル中の名詞で2012年から2013年にかけて最も増加度合いの高い表現は「推定」

という結果が得られた。図1に示すとおり、タイトル中に「推定」という表現を含む論文の件数は基本的に増加傾向にある。比較対象として、データ全体の傾向を図 2 に示す。

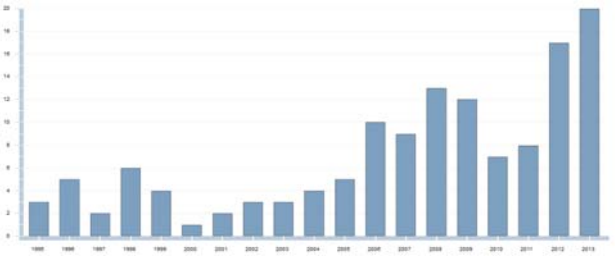

図 1: タイトル中に「推定」という表現を含む論文の年別件数

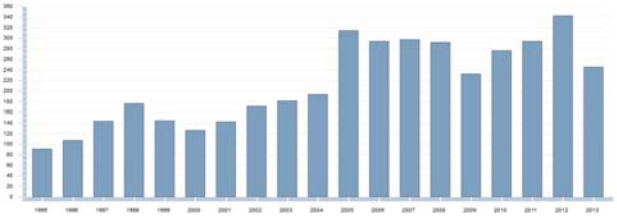

図 2: 全データに含まれる論文の年別件数

4.5. 処理データ量の変遷の分析

他に何か面白い現象を分析できないかという観点から、論文中に出現する数量を調べ年別にプロットした。

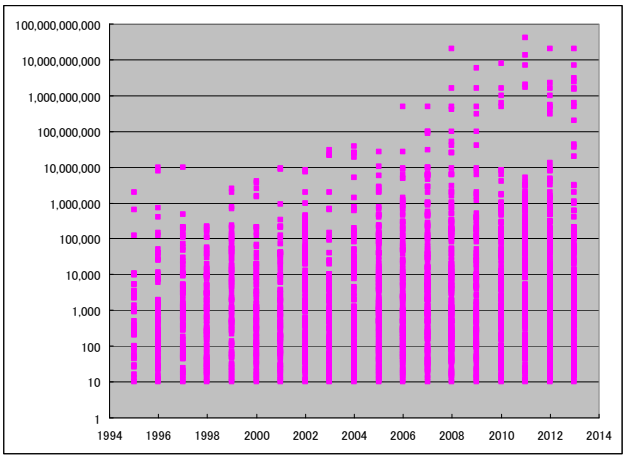

図 3: 論文中に出現する数量の年別傾向

図 3 では、「文」「文書」「件」「単語」「文字」の 5 種類の単位について、論文中に出現した数字をプロットしている。論文中の出現回数や出現論文件数は考慮せず、1年に1回でも出現した場合にプロットしている。縦軸は対数軸であり、単に数値を出しているため、様々な単位の数値が混在しているが、基本的な傾向として、新しくなるほど規模が大きくなってくる様子が示されている。

4.6. メタデータ抽出と抽出結果を用いた分析

テキストマイニングの難しさは言語表現の曖昧性にある。多義性などの理由により、ある表現がテキスト中に存在しても、その解釈が一通りには決まらないことが多い。そこがテキストマイニングとデータマイニングの大きな違いであり、安定した解釈が可能なメタデータ(今回対象としたデータの場合は、収録論文一覧表から得られる発表年や著者名など)をうまく活用することが、信頼性の高いマイニング結果を得る重要なポイントとなる。

今回対象としたデータの場合、収録論文一覧表中のメタデータ以外に、本文中「科研費」か「科学研究費」という表現を含む論文は、科学研究費助成事業の助成を受けた研究の成果であることが比較的確実な情報として特定できることが認められた。こういった情報を活用することで、例えば科学研究費助成事業を活用した研究の特徴分析が容易になる。

5. おわりに

19 年分の年次大会予稿集データを知的資産として活用する可能性を検討し、テキストマイニング技術を用いて実データの分析を行い、その妥当性を確認した。

言語処理学会大会予稿集データの分析に関しては、村田らによる第 10 回までの 10 年分の動向調査[5]が存在する。本稿では、対象期間が長くなっただけでなく、単なる動向調査という範囲を超え、19 年分の文献が知的資産としてどう役立つかを検討し、多様な活用目的の妥当性を実際に試行して確認した。また、村田らの調査では書誌情報のみを対象とし、自然言語処理はタイトルを形態素解析する程度の比較的浅いレベルにとどまっていたのに対し、本稿では本文全体を解析し活用した。

4.1.で示した「技術の特徴の分析」は、福田ら[6]の「特定分野の技術文献からの要素技術とその効果を示す表現の抽出」に近いが、自然言語処理という特定分野の技術文献において、比較的単純な仕組みでも特定技術に特徴的な表現を得られることを確認できた。

今回対象とした約 4 千というデータの件数は、本格的なテキストマイニングの対象としては、極めて小規模であり、人手である程度把握することのできる量である。そのため、人手では気付かず、何らかのアクションにつながるような知見の獲得は期待できない。結果的に本稿の取組みで得られた知見は、長年この分野で研究を続けている人間にとっては意外性の無い内容である。しかし、初学者や分野の異なる研究者にとっては有用な知識が得られる可能性があり、その観点から、この対象データを知的資産として活用することを検討した。

今後のあるべき姿として、まずは本稿で示したような活用目的でデータの蓄積と利用環境の整備を進め、整備が進んだ段階で、他分野も含めた大規模な量の技術文献をテキストマイニングできる環境を構築することが望ましいと考える。それによって、例えば、言語処理分野の文献を扱っているだけでは気付かなかった自然言語処理の特徴や、他分野の技術との融合の可能性など、有用な知見を得られる可能性が生まれてくると考える。

参考文献

[1] Marti A. Hearst. Untangling Text Data Mining. In Proceedings of ACL-99, pp. 3-10. 1999

[2] 那須川哲哉. テキストマイニングを使う技術/作る技術—基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006

[3] Wei-Dong (Jackie) Zhu, Asako Iwai, Todd Leyba, Josemina Magdalen, Kristin McNeil, Tetsuya Nasukawa, Nitaben (Nita) Patel, and Kei Sugano. IBM Content Analytics Version 2.2: Discovering Actionable Insight from Your Content. ISBN: 0738435287. 2011 http://www.redbooks.ibm.com/abstracts/sg247877.html

[4] 西山莉紗, 竹内広宜, 渡辺日出雄, 那須川哲哉. 新技術が持つ特長に注目した技術調査支援ツール, 人工知能学会論文誌, Vol. 24, No. 6, pp.541-548. 2009

[5] 村田真樹, 一井康二, 馬青, 白土保, 井佐原均. 過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査. 言語処理学会 第 11 回年次大会, 2005

[6] 福田悟志, 難波英嗣, 竹澤寿幸. 論文と特許からの技術動向情報の抽出と可視化. 情報処理学会論文誌. データベース 6(2), 16-29. 2013