テキストマイニングシンポジウムでの発表内容と言語処理技術

竹内 孔一 [1]
岡山大学大学院

金山 博 [2]
日本アイ・ビー・エム株式会社 東京基礎研究所

市瀬 眞 [3]
株式会社 NTT ドコモ 情報システム部

榊 剛史 [4]
株式会社ホットリンク

渡辺 靖彦 [5]
龍谷大学 理工学部

東中竜一郎 [6]
日本電信電話株式会社 NTT メディアインテリジェンス研究所

嶋田 和孝 [7]
九州工業大学 大学院情報工学研究院

[1]koichi@cl.cs.okayama-u.ac.jp, [2]hkana@jp.ibm.com, [3]ichisem@nttdocomo.com,
[4]t.sakaki@hottolink.co.jp, [5]watanabe@rins.ryukoku.ac.jp,
[6]higashinaka.ryuichiro@lab.ntt.co.jp, [7]shimada@pluto.ai.kyutech.ac.jp

1 はじめに

電子情報通信学会 言語理解とコミュニケーション研究会[1]では，2011 年からテキストマイニング・シンポジウムを開催しており，2016 年 2 月で 8 回を数えた．これは学術界からの研究成果と，産業界での実践的な知見に基づく技術や，実務に使う側の知見や要望を合わせて議論する場として定着してきている．本稿では，過去 5 年のシンポジウムの発表から，学術側から見て特徴的なものを取り上げ，議論されてきたテーマ，提案された技術，未解決の課題などについて論じたい．また事例を取り上げた後，テキストマイニング全てに共通した言語処理の置かれている位置付けを確認し，実社会の要求に応える言語処理の可能性について議論する．

2 テキストマイニングの目的と基本的な課題

テキストマイニングには，第 2 回シンポジウムの那須川氏の講演 [6] にあるように,「大量のテキストデータから役立つ知見を得る」，より具体的には「個々のテキストの情報だけからは得られない知見を得る」[6] という目的があると考えられる．また筆者が考えるテキストマイニングの特徴は，この目的を達成する状況として「何を取りだして良いか分からない」という状況からスタートことがあり，検索のタスクでは本質的に解決できない点である．例えば企業のコールセンターに蓄積されたテキストの中で，何が問題になっているかは，キーワードを集約するだけでは把握することが難しいが，人手で個々のテキストを全て読んで整理することもまた量的に不可能である．文献 [6] が指摘するように，クラスタリングで抽象化すると意図が不明になってしまい，文字面ベースだと表現の異なりで分散してしまう．取り出したい情報が不明な場合，異なる表現を同一視するための辞書を予め作成することは人手でも不可能である．これに対して [6] では，単語より長い単位 (X が V できない) での表現の集約を行いつつ，あらゆる語 (または商品名など分野に特化した語句) またはフレーズなどと数値的な比較をすることで実際に有益な知見を得る方法を実践している (例えば文献 [4])．この状況から，テキストマイニングという研究分野に下記の 2 点の特徴を見いだせる．

1 主眼は有益な知見 (とそのエビデンス) の獲得であり，ツールではない

2 知見を得るためには，ツールを用いて操作する作業者の知識も求められる

従って，言語処理技術の精度の改善が，テキストマイニングの効果に直接的に反映されるとは限らないというのが現実である．しかし，分野依存辞書の構築 [7]

[1]http://www.ieice.org/~nlc/

など共通の課題は存在するのも確かである．以下，実データに対してどのような要求があるか，どのような分析が行われてきたかを提示することで，現実のタスクに直結するような新たな言語処理研究の課題の創出に貢献したい．

3 テキストマイニングで発表された内容

シンポジウムでは学術的な発表，企業デモ，討論などさまざまな発表スタイルを設けている．その中で本稿では学術的な要素を含みつつ，現実の問題に対して研究を行っている例を紹介する．これによってどんな課題でどういう情報を取り出す必要があるか，また取り出したものが社会的にどういう価値があったかを示すことで実社会に必要とされる言語処理への事例を提示したい．

1) 企業の業績・活動に対するテキストマイニング

記事や SNS から企業活動について企業情報を収集して有益な情報を獲得しようとする研究報告が 10 件以上報告されている．その中で経済動向や株価推定の研究が発表されている．和泉ら [3] は日本銀行の金融経済月報を利用して，月ごとの単語の主成分スコアの時系列を特徴として，回帰分析を当てはめることにより翌月の日本国債市場の運用をテスト評価として行った．その結果，テキストを利用したときの方が他の数値を利用した予測より高い利益を得ることを実験的に示した．

羽室ら [2] は投資家が近年の配信される金融関係の評判テキストに左右されているかどうか分析するために，Bloomberg 社の記事に含まれる評判情報 (「需要が伸びる」や「株価が反発する」など) が株価変動にどのように影響を与えているかを分析している．ここで企業の評判情報を獲得するための評価表現辞書の構築のために，那須川ら [7] の手法を利用している．これにより「景気が回復する」という格助詞と用言のペアによる辞書を構築している．評価表現辞書を利用して，記事からセンチメント指数を求め，株価との相関を調べたところ高い相関があることを示した．また，シミュレーションによる運用実験でセンチメント指数を入れた場合に実用的に有効であることを示した．

薄井ら [8] も企業活動ニュースにおける評判評価情報に着目したが，さらに表現を細分類してニュースのセンチメント値を求める手法を提案している．まず評価辞書の構築としてニュース記事に対して形態素を tf-idf により重み付けして重要語のみを抽出する (これをキーワードと呼ぶ)．次に，各キーワードの極性についてはキーワードを含むニュースが配信されたとき，株価が上昇したか下降したかで極性を判定し，重回帰分析を用いて評価値を付与する．この方法により例えば「業務改善命令」や「下方修正」など企業活動の評価に必要な語が獲得できている．これを高村らが作成した極性辞書 [1] と比較したところ，高村らの辞書はこれらのうちの 2.6%程度しか網羅していないことがわかった．このキーワードベースの評価辞書を用いて，ニュース記事の極性を判定する．その際，単にキーワードを含む場合の文と,「売り上げ減少に伴い，赤字に転落した」といった原因-結果を含む評価文を別に評価した．これは因果関係は株価の影響に対して大きいと考えられるためためニュース評価の際により大きな重みを与えるためである．こうして作成したニュース記事センチメント分析手法を 1000 文のニュース記事と配信後の株価の値動きで評価したところ，プラス評価に対して 7 割の一致率，マイナス評価に対して 4 割の一致率を得たことを示している．精度としてはまだ低いが，ニュース配信後の株価をテキストに対する評価として利用している部分が興味深い．

また杉原らは営業日報から課題文を取り出し，顧客との商談の可能性を広げる取り組みを行っている．また坂地，中山，西沢は企業活動の指標を取り出すことで，情報集約を行うなどしている．こうした研究はテキストからの情報抽出に近く，抽出すべきものが明確で有り，それらと実社会での企業活動や価値との相関を明らかにしている．一方で，大森は数年にわたる電機業界の活動に対してテキストマイニングを行い，成長している企業とそうでない企業との差について海外との標準化や研究への投資があることを明らかにした．業績そのものは数値であるが，要因はテキストにしかなく，マイニングツールを利用した分析による事例を提供している．酒井 [5] らは企業活動と就職活動時のキーワードがマッチしていないことに気づき，企業の業績発表記事から活動を表す適切なキーワードを抽出する手法を提案している．

2) 医療介護福祉

医療や介護に関する発表は 4 件あり，実務的な課題を明らかにしている．山下らは病院における長期在院者の特徴を推定する研究を提案した．また福田らは介護現場にテキストマイニングを適用して，申し送り情報の中で，取りこぼされていたモノを単語の共起グラフから獲得して実際の改善に繋いでいる．

3) 政策にかかわる意見集約

政策に関する研究では木村らは地方の政治会議議事録から政策として重要な案件がなにかを集約する方法について試みている．また岩見らはエネルギー政策に関するパブリックコメントから意見を集約するために特徴的な議論を可視化ツールを利用しつつまとめて，意見の分析結果を報告した．

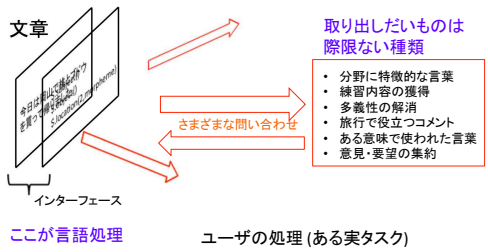

図 1: 言語処理は文書に対するあらゆる問い合わせを受けるインターフェース

こうした上記のような大きなテーマの他に高齢者が開いた時間に個人のスキルをいかして働けるようにスキルマッチング手法を考察する研究や，テキスト記事から未来予測部分を取り出すことで未来予想を取り出す手法などが提案されている．

4 言語処理の位置付けと発展

上記で取り上げたテキストマイニングに関する研究は処理に主眼があるわけではなく，取り出したテキストに価値があることが明らかである．一方で，言語処理はテキストから必要な情報 (それは分野依存であり，前もって分野非依存で用意することが不可能な情報) を取り出すためのありとあらゆるテキストに対する操作が要求される部分であるということである．例えば，企業情報であれば,「企業活動を表す文書」を集める必要があり，文の中では「企業名」やその「活動 」表現する部分を獲得し，表現の正規化が必要になる．それらをこなすツールは存在しないため,「企業活動を表す文書」を表すには，そうした記事を書いてるニュースサイトを固定したり,「活動」などはキーワードを決めるか「動詞」といった品詞レベルで押さえるといった手法しかない．よって問題・分野に依存した，テキスト情報抽出手法の開発は有益である．さらに，テキストマイニングの主は価値ある情報であり，分析者はツール構築に興味は無い．この部分において，言語処理を研究している研究者が分析者と共同で活動することでより具体的な実処理に役立つ研究テーマと成果が得られるのではないかと推察できる．

参考文献

[1] H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Cumputational Linguistics*, pp. 133–140, 2005.

[2] 行信羽室, 克彦岡田. テキストマイニングを用いた株式銘柄センチメントの測定とポートフォリオの構築 : マーケット・ニュートラルアプローチ. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 113–118, 2011.

[3] 和泉潔, 後藤卓, 松井藤五郎. 経済リポートのテキスト分析による金融市場動向推定. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 107–111, 2011.

[4] 竹内広宜, 那須川哲哉, 渡辺日出雄. コールセンターにおけるビジネス会話のマイニング. 人工知能学会論文誌, Vol. 23, No. 6, pp. 384–391, 2008.

[5] 酒井浩之, 坂地泰紀. 企業 web ページを対象とした企業検索システムのための検索クエリに関連するタグの推定. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 41–45, 2014.

[6] 那須川哲哉. テキストマイニングの可能性〜有用性と研究の発展性〜. 言語理解とコミュニケーション研究会基調講演, 2012.

[7] 那須川哲哉, 金山博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会第 162 回自然言語処理研究会報告, pp. 109–116, 2004.

[8] 駿希薄井, 博哉吉田. ニュース記事を用いたセンチメント分析に基づく企業評価システムの開発. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 1–4, 2014.