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
岡山大学大学院
koichi@cl.cs.okayama-u.ac.jp

金山 博
日本アイ・ビー・エム株式会社 東京基礎研究所
hkana@jp.ibm.com

市瀬 眞
株式会社 NTT ドコモ 情報システム部
ichisem@nttdocomo.com

榊 剛史
株式会社ホットリンク
t.sakaki@hottolink.co.jp

渡辺 靖彦
龍谷大学 理工学部
watanabe@rins.ryukoku.ac.jp

東中竜一郎
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
higashinaka.ryuichiro@lab.ntt.co.jp

嶋田 和孝
九州工業大学 大学院情報工学研究院
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．この会議は，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，実務で利用する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 年間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 [18] 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」[18] 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することがあり，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 [18] 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 [18] 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 [9])．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 [19] 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・シンポジウムで発表された内容

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

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

以上報告されている．その中で特に目立ったのは経済動向や株価推定の研究である．和泉ら [7] は日本銀行の金融経済月報を利用し，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 [11] 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 [19] の極性辞書構築手法**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，過去のデータに対す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 [14] 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に対して影響が大きいと考えられる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．

廣川ら [12] は医薬品製造業 68 社を対象に有価証券報告書のテキストから特徴語を抽出し，単語ベクトルに基づく SVM を適用することで，当該期の企業の利益伸び率が上位 α 位に入ってるかどうかを判定する手法を提案した．これらテキストを利用した企業活動推定のポイントは株価や利益率など測定できる数値が存在し，なおかつ発行時期が明確な文書が存在することにある．よってこうした良質なデータが存在すれば企業活動の予測が可能であることがうかがえる．

一方，こうした評価値との結びつきとは異なる研究も報告されている．杉原ら [17] は営業支援システムに蓄積される営業日報テキストデータから課題記述文を取り出し，顧客との商談の可能性を広げる取り組みを行っている．課題記述文とは「望ましくない状況や望ましいゴールといった解決・改善の対象や結果として記述される文」であり，改善策や問題点が記述されている文である．例えば「人力作業が多く，それに伴う工数やミスを減らしたい」といった文を取り出す．**SVM による抽出モデル**を仮定し，特徴量として課題文に現れやすい，トラブルや要求，解法や評価表現を取り出すための単語を指定するために**言語資源**を利用する[2]．さらに，文書内での文の位置などの文の特徴，自立語 n-gram，極性語と PMI 値の高い語の頻度数を特徴量とした．その結果 F 値で 40%程度の精度を得たことを報告している．また酒井ら [10] 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．

これら上記の研究はいずれもテキストから抽出すべきものが比較的明確であるため情報抽出に近いタスクと考えられる．一方で，大森 [3] は数年にわたる電機業界の活動に対して成長要因を分析するテキストマイニングの結果を報告している．この際，知識の構造化手法を取り入れた独自の分析フレームワークを仮定し，テキストから得られた**単語共起グラフ**の解釈から，テキストと企業活動の指標を参考に，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．こうした数値指標とテキストを元にした要因分析は分析者の知識構造に頼る部分が多く，自動化できる部分がほとんどないことがこの研究から分かる[3]．

2) 医療・介護・福祉関連

医療や介護に関する発表が数件あり，実務的な課題を明らかにしている．山下ら [5] は病院における手術後の在院日数に着目して，長期在院者の特徴を推定する研究を提案した．診療データなどのテキストデータから在院日数に影響あたえた要因は何かを取り出すのが目的である．手法としては手術記録文書から医学辞書を利用して重要語を抽出し，**SVM** を利用して 25 日以上在院した場合を正例として，正例に寄与した単語を収集するものである．獲得できた単語が長期滞在にどのように関連していたかはさらなる分析の必要があるが因子分析の可能性を示した．

大山ら [6] は介護施設に対してアンケートを行い，若年性認知症患者の受け入れ拒否理由について得られた自由回答文に対してテキストマイニングを行った分析を発表している．**単語間の共起グラフ**から「トラブル」と「暴言」「暴力」「ケンカ」との共起が高いことがわかり，実際のアンケート文を確認したところ，施設側がトラブル発生時の対処に懸念を抱いてることが要因であることを明らかにした．

---

[2] 日本語評価極性辞書 (http://www.cl.ecei.tohoku.ac.jp/index.php?Open Resources/
Japanese Sentiment Polarity Dictionary)，および「負担・トラブル表現リスト」(https://alaginrc.nict.go.jp/)

[3] この研究は言語理解とコミュニケーション研究会 2013 年研究賞を受賞．

福田ら [8] は介護現場における申し送り情報に対して**単語間の共起グラフ**に基づくテキストマイニングを行い，業務改善した実例を複数報告している．1 例をあげると，共起グラフは通常，介助に関する言葉が現れるのに対して，「夫」，「差し入れ」，「黙る」など異なった共起語が現れた．これをもとに職員で振り返ったところ，利用者の夫が介護スタッフの見ていないところで利用者である妻に食事を差し入れていることが判明した．利用者は飲み込みが弱く誤嚥の可能性があるため，対処として職員のいるところで食事を与えることを認めたところ，利用者のご家族の満足度が大きく向上した[4]．

こうした医療まわりの事例からわかることは，人のケアに関わる部分は些細なことでも当該者にとって重要なできごとであり，個別の対応が求められることが想像される．よって抽象化や数値化といった全体の傾向を分析するというよりテキストをベースにどのようなケアが必要かを取り出すことが優先される課題分野と考えられる．現段階では単語の共起グラフから読み取る以上の手法が見受けられないが，テキストマイニングが活用されるべき課題と考えられる．

3) 政策にかかわる意見集約

葦原ら [13] は地方の議会議事録から政策として求められている要望・要件を取り出す手法を提案している．議会議事録は通常のテキストと異なり，議員の質問と回答など，会話になっていること，また，一文が長く並列構造が多用されるなどの特徴がある．そこで**CBAP[2] を利用した文節単位での処理**を提案し，要求を表す末尾表現である「べき」がどの文節にまで影響するか，議事録を分析して特徴量とした．要求部分の抽出には **SVM** を利用し，ベースライン (Cabocha+日本語機能表現辞書) に対して F 値で 4 ポイント以上高いことを示した．

また岩見ら [20] はエネルギー政策に関するパブリックコメント 9 万件を可視化した結果について報告した．手法としては意見を人手により特徴的であると考えられるクラスに分類し，特徴クラス間の**ネットワークグラフ**を描画して意見の構造化を試みた．しかしながらネットワーク構造が複雑で有り，ネットワークから全体的な構造をどう理解するかについては明確な結論は得られなかった．

このように政策に関する意見収集は表現の自由度が高く，数値化の見込みも低いためテキスト表現が主となるが，単純な評価文ではないため明確な分析手法が見えていない状況である．既存の係り受け解析だけでなく，文節単位の処理など，長い文に対する処理を強化した言語処理システムが求められる．

4) その他

上記のような大きなテーマの他に高齢者が空いた時間に個人のスキルを活かして働けるようにするスキルマッチング手法 [4] 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 [16]，陸上競技におけるライバルの活動状況を獲得してモチベーションを向上する手法 [15] など具体的でテキストに埋もれている有益な情報を利用しようとするアイデアにあふれた研究が提案されている．テキストから価値ある情報を取り出す分野が広く，またテキストが同じ性質でないため各問題に応じた言語処理ツールが求められる．

[4] この研究は言語理解とコミュニケーション研究会 2014 年研究賞を受賞．

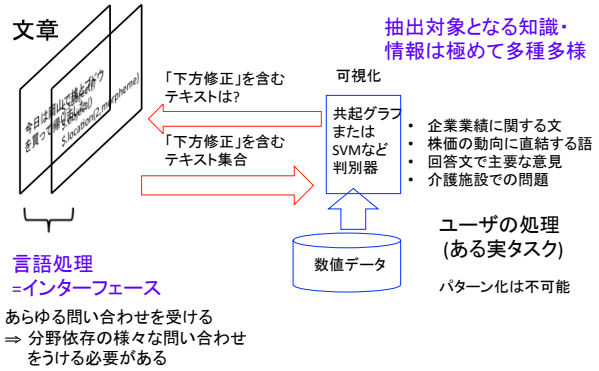

図 1: 言語処理は文書に対するあらゆる問い合わせを受けるインターフェース

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で利用された言語処理や手法など太字でマークした．その結果，テキストを特定のキーワード (極性辞書や分野依存の表現) の共起グラフがよく使われていることがわかる．また評価値 (株価などの評価指標) がある場合は SVM が使われている (この場合はキーワードは特徴量として利用される)．どちらの場合も，“テキストに対する操作” としての言語処理を考えるとキーワードマッチングが主であり，キーワード単独の頻度や共起頻度などをある限られたフィルタ (名詞のみなど) で獲得する操作が中心となる．形態素解析も係り受け解析も，キーワードが目的とする使われ方をしているかどうかを指定するためのパターンを記述するために利用される．

これは例えばテキストから取り出したい内容が企業情報であれば，「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」を表現する部分を獲得し，表現の正規化が必要になる．しかし直接こうした文を獲得する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い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．

つまりテキストから必要な情報を取り出す状況において，分野非依存のツールだけでは解決せず，問題・分野に依存したテキスト情報抽出手法を分析者側が構築できないためキーワードベースでの方法で代替している状況であると考えられる．この状況を図 1 に示した．よって言語処理はテキストに対してあらゆる要求

に対して情報を獲得できるツールを構築していく必要があるのではないかと考えられる．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思われる．テキストマイニング・シンポジウムでは，引続きこの点を遡及していきたい．

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] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 節境界自動検出ルールの作成と評価. 言語処理学会第九回年次大会発表論文集, pp. 517–520, 2003.

[3] 大森寛文. 電機業界における経営課題の認識構造と実行動に関する知識の発見 : 知識の構造化論とテキストマイニングの融合. 電子情報通信学会技術研究報告. 113:213 (NLC), 第 3 回テキストマイニング・シンポジウム, pp. 83–88, 2013.

[4] 三浦貴大, 小林正朋, 檜山敦, 高木啓伸, 廣瀬通孝. 高齢者の履歴書からの特徴語抽出によるスキルの発見とマッチング. 電子情報通信学会技術研究報告. 112:196 (NLC), 第 2 回テキストマイニング・シンポジウム, 2015.

[5] 山下貴範, 若田好史, 濱井敏, 中島康晴, 岩本幸英, フラナガンブレンダン, 中島直樹, 廣川佐千男. 手術記録から術後在院日数を特徴付ける重要因子抽出モデルの構築. 電子情報通信学会技術研究報告. 114:211 (NLC), 第 5 回テキストマイニング・シンポジウム, pp. 1–6, 2014.

[6] 大山恭史, 池田望. テキストマイニングによる介護施設の利用者受入要因の分析 : 若年性認知症患者の受入調査から. 電子情報通信学会技術研究報告. 114:211 (NLC), 第 5 回テキストマイニング・シンポジウム, pp. 7–9, 2014.

[7] 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. 111:119 (NLC), 第 1 回テキストマイニング・シンポジウム, pp. 107–111, 2011.

[8] 福田賢一郎, 濱崎雅弘, 福原知宏, 藤井亮嗣, 堀田美晴, 西村拓一. 介護現場における申し送り情報の分析 : 業務改善に向けて. 電子情報通信学会技術研究報告. 114:211 (NLC), 第 5 回テキストマイニング・シンポジウム, pp. 11–16, 2014.

[9] 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.

[10] 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. 114:211 (NLC), 第 5 回テキストマイニング・シンポジウム, pp. 41–45, 2014.

[11] 羽室行信, 岡田克彦. 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. 111:119 (NLC), 第 1 回テキストマイニング・シンポジウム, pp. 113–118, 2011.

[12] 廣川佐千男. 文単位の有価証券報告書分析による利益伸び率の予測. 電子情報通信学会技術研究報告. 113:213 (NLC), 第 3 回テキストマイニング・シンポジウム, pp. 77–82, 2013.

[13] 葦原史敏, 木村泰知, 荒木健治. 節の分類情報を用いた地方議会会議録における要求・要望表現抽出. 電子情報通信学会技術研究報告. 112:196 (NLC), 第 2 回テキストマイニング・シンポジウム, pp. 1–6, 2012.

[14] 薄井駿希, 吉田博哉. 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. 113:429 (NLC), 第 4 回テキストマイニング・シンポジウム, pp. 1–4, 2014.

[15] 佐野正和, 福原知宏, 増田英孝, 山田剛一. 陸上競技ブログからの活動記録抽出の試み. 電子情報通信学会技術研究報告. 115:445 (NLC), 第 8 回テキストマイニング・シンポジウム, 2016.

[16] 島岡聖世, 佐藤祥多, 佐々木彬, 稲田和明, 関根聡, 乾健太郎. 条件付き確率場を用いた新聞報道からの未来予測情報抽出. 電子情報通信学会技術研究報告. 115:222 (NLC), 第 7 回テキストマイニング・シンポジウム, 2015.

[17] 杉原大悟, 大熊智子, 佐竹功次, 三浦康秀, 服部圭悟, 増市博. 営業支援システム内に蓄積されたテキストデータからの課題記述文抽出. 電子情報通信学会技術研究報告. 112:196 (NLC), 第 2 回テキストマイニング・シンポジウム, pp. 7–12, 2012.

[18] 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〜. 電子情報通信学会技術研究報告. 112:196 (NLC), 第 2 回テキストマイニング・シンポジウム, 2012.

[19] 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.

[20] 岩見麻子, 木村道徳, 松井孝典, 熊澤輝一. 大規模パブリックコメントの意見構造の把握と可視化〜「エネルギー・環境に関する選択肢に対する御意見の募集」を事例として. 電子情報通信学会技術研究報告. 115:445 (NLC), 第 8 回テキストマイニング・シンポジウム, 2016.