応用

3.12 Web 応用タスクにおけるエラー分析

― Twitter を用いた疾患サーベイランスを題材に―

荒牧 英治（奈良先端科学技術大学院大学）　岡崎 直観（東北大学）

タスクの定義

2000 年以降の自然言語処理（NLP）の発展の一翼を担ったのは World Wide Web（WWW）である．Web を大規模テキストコーパスと見なし，そこから知識や統計量を抽出することで，形態素解析，構文解析，固有表現抽出，述語項構造解析，機械翻訳など，さまざまなタスクで精度の向上が報告されている．これらは，Web が NLP を高度化した事例と言える．

同時に，Web を社会の活動を記録したビッグデータと見なし，それを分析することで，日常生活やビジネスでの意思決定に活かそうという試みも盛んに取り組まれている．データ分析の 3 段階[1]，すなわち現象分析的な分析（descriptive analysis），未来予測的な分析（predictive analysis），戦略指示的な分析（prescriptive analysis）の中で基本になるのは，Web のデータから過去および現在の状況・現象を理解する現象分析的な分析である．Web のテキストデータから過去および現在の状況・現象を理解することは，Web というメディアを通して発信されたテキストから世の中の出来事を正確に復元することである．Twitter や Facebook などのソーシャルメディアの分析では，個人による情報発信や拡散性，即時性，双方向性などの特徴が加わり，地震震源地の特定[2]，災害情報の整理☆1，感染症のサーベイランス☆2 などの新しい応用が生まれている．

本プロジェクトでは，Web のテキストデータから個人の実際の経験や意図を推測する（マーケティングでは，「傾聴」という言葉が用いられている）というタスクにおいて，自然言語処理の最先端技術の適用と，そのエラーの分析，取り組むべき課題の整理を行った．

☆1 http://www.nict.go.jp/univ-com/isp/research.html
☆2 http://mednlp.jp/influ/

技術の紹介

具体的には，さまざまな自然言語処理技術をさまざまな Web 文章に適応することで，現実世界の情報をリアルタイムに可視化する．これには，単なる言語処理技術を超えた他分野の技術も必要となる．たとえば，大量 Web テキストをリアルタイムに処理するデータベースに関する技術や，可視化サービスを行うユーザインタフェースに関する技術も必要となる．この応用指向が本質的な研究を困難にする場合もある．たとえば，大規模な Web データに対して自然言語処理技術を適用し，社会の動向を迅速かつ大規模に把握しようという取り組みは，対象とするデータの性質に強く依拠する．そのため，より一般的なほかの自然言語処理課題に転用できる知見や要素技術を抽出することが難しいという課題もある．

これらを踏まえた上で次の 2 点が Web 応用の本質的な課題であると考えている．

(1) NLP 課題の明確化：ソーシャルメディア上のテキストの蓄積を自然言語処理の方法論で分析し，人々の行動，意見，感情，状況を把握しようとするとき，現状の自然言語処理技術が抱えている問題を認識すること
(2) 共通タスクの切り出し：応用事例（たとえば疾患状況把握）の誤り事例の分析から，自然言語処理で解くべき一般的な（複数の応用事例にまたがって適用できる）課題を整理すること

これらについて，本タスクは，風邪とインフルエンザの流行把握を題材にして取り組んだ．このタス

誤り分類	説明	事例	誤り事例数（割合）
非当事者	疾患・症状を所有する対象が，発言者およびその周辺の人物ではない場合	みんな風邪ひかないように暖かくして寝ようね！	100（23.5%）
比喩	比喩的に疾患表現が利用されている場合が当てはまる	凄すぎて鼻水ふいた ww	87（20.4%）
一般論	そもそも疾患・症状の保有に関する話題ではなく，疾患そのものについて議論している場合	風邪ウイルスが目に見えたらなあ	63（14.8%）
モダリティ	「かもしれない」（疑い），「かな？」（疑問）などのモダリティ表現により，疾患の事実が認められない場合	風邪をひいたときはお肉を食べましょう	46（10.8%）
時制	疾患のあった時間が異なる場合	高熱で夜中中うなされても次の日出勤できるから助かる	43（10.1%）
否定	疾患の事実が否定されている場合	ノドいたた．風邪のようでなんかちがう，なんじゃろ	25（5.9%）
その他	その他	—	62（14.6%）

表 -1　誤り分類と，その割合

クは，Twitter 上での発言者が該当する疾患を持つかどうかを判定するタスクである．これは，文章分類タスクの一種と考えられ，単語 n-gram 素性を用い Support Vector Machine（SVM）にて分類を行う手法が提案されている [3]．この誤りを分析した結果，**表 -1** のような結果となった．

これらの事例は，上記 6 つに大別されるが，言語処理の研究課題という観点から整理すると，疾患があったのかという事実性（時制，モダリティ，否定，比喩）と，仮に疾患の事実があったとして，疾患を所有しているのは誰なのかという主体性（非当事者や一般論の問題）という 2 つの大きな言語現象に大別できる．これらの判定精度をいかに向上させるかが Web 応用の本質的な課題となる．

近い将来の達成可能性

今後，事実性解析および主体性解析について，Web 応用の実用面から研究が活発化すると考えられる．事実，本タスクをベースにし事実性解析 [4] と主体性解析 [5] に関する研究が本年度発表された．本来，自然言語処理は，特定の言語やタスクを仮定しない研究分野ではあるが，Web テキストを扱う以上，どのように使うのかといった応用面を考慮するのが自然である．Web 応用は，基礎から応用までをカバーしたやりがいのあるタスクであり，今後も，NLP の発展の一翼を担う技術である．

参考文献

1) James, R. E. : Business Analytics : The Next Frontier for Decision Sciences, Decision Line, 43(2), pp4-6 (2012).
2) Sakaki, T., Okazaki, M. and Matsuo, Y. : Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors, The 19th International Conference on World Wide Web (WWW), pp.851-60 (2010).
3) Aramaki, E., Masukawa, S. and Morita, M. : Twitter Catches The Flu : Detecting Influenza Epidemics Using Twitter, EMNLP2011, pp.1568-1576.
4) Kitagawa. Y., Komachi, M., Aramaki, E., Okazaki, N. and Ishikawa, H. : Disease Event Detection based on Deep Modality Analysis, ACL-IJCNLP Student Research Workshop, pp.28-34 (2015).
5) Kanouchi, S., Okazaki, N., Komachi, M., Aramaki, E. and Ishikawa, H. : Editors. Who Caught a Cold ? Identifying The Subject Who has a Symptom, ACL-IJCNLP, pp.1660-1670 (2015).

（2015 年 10 月 15 日受付）

荒牧 英治（正会員）aramaki@is.naist.jp
2000 年京都大学総合人間学部卒業．2005 年東京大学大学院情報理工系研究科博士課程修了．博士（情報理工学）．以降，東京大学医学部附属病院特任助教を経て，奈良先端科学技術大学院大学特任准教授．医療情報学，自然言語処理の研究に従事．

岡崎 直観（正会員）okazaki@ecei.tohoku.ac.jp
2007 年東京大学大学院情報理工学研究科博士課程修了．同研究科・特別研究員を経て，2011 年より東北大学大学院情報科学研究科准教授．自然言語処理，テキストマイニングの研究に従事．