統計数理(2016)
第 64 巻 第 2 号 141–143

「特集 統計的言語研究の現在」編集にあたって

持橋 大地[1]・前川 喜久雄[2]・浅原 正幸[2]

統計学と言語学の間には過去にも若干の交流があった．本誌のバックナンバーを繰ってみると，26 巻 1–2 号(1979 年刊)に村上征勝氏が「著者推定問題における統計的手法」と題した研究ノートを寄稿しておられるのが見つかる．その後 20 年ほどの間隔をあけて，48 巻 2 号(2000 年刊)に同じ村上氏がエディターを務めた「“ことば” 新研究」という特集が掲載されている．そのまえがきには，シェークスピア＝ベーコン説などの著者推定問題に始まり，その後細々と続いてきた言語に関する統計的研究が，「ここに来て，大きなうねりとなる兆候が見えてきている．というのも近年のコンピュータの著しい進歩により，大量の “ことば” のデータベースが構築できるようになり，加えて，安価で高性能のパソコンの普及により，比較的簡単に複雑な統計分析ができるようになった為である．“ことば” の分析は新たな段階に入ったといえる.」との記述がある．

村上氏が 20 世紀末に指摘したこの「うねり」は，その後，インターネットがもたらすビッグデータの登場とも呼応して，現在では関連学会のみならず産業界をも飲みこむ大波に成長している．現在，音声認識や自然言語処理の主流が機械学習と連携した統計的なアプローチにあることは，広く認識されているとおりである．

20 世紀末と現在を比較してもうひとつ気づくのは，統計手法の利用目的が仮説検定や主成分分析といった事後の分析から，データのモデリングに移行していることである．一般化線形混合モデル，MCMC による階層ベイズモデルなどの普及により，統計ユーザーがみずからの発意で自由に統計モデルを構築できるようになったことは，言語分析にかぎらず，人文学領域の問題に対する統計学の応用可能性を飛躍的に高めたように思われる(その背後には R や Python に代表される，無償かつ高機能な計算環境の普及があることはいうまでもない)．

本特集の出発点となったのは，「統計的言語研究の現在」と題された国立国語研究所と統計数理研究所の合同シンポジウム(2015 年 9 月 4 日開催)であった．国立国語研究所講堂で開催されたこのシンポジウムは，われわれの予想をこえる参加者 116 名の盛会となり，言語研究における統計的手法の注目度の高さを感じさせるものであった．シンポジウムの内容，および講演スライドはすべて，ホームページ[1)]から今でもご覧いただくことができる．本特集所載の論文のうち，新井論文，村脇論文，荒牧論文はこのシンポジウムでの講演者の手になるものであり，他の論文も，シンポジウムの企画段階で講師の候補に名前があがった方々に執筆していただいている．他に言語心理学や社会言語学の領域からも寄稿していただく予定があったが，諸般の事情で実現に到らなかったのは残念であった．

従前の統計的言語研究は，主に単語などの頻度情報に基づくものであった．一方，本特集で扱う言語のデータは多様な形式からなる．単語列や品詞列などを抽象化した系列，系統樹や構文構造の根幹をなす木構造，被験者の反応・経年変化の手がかりとなる時間，さらには Twitter 発言の GPS 情報や言語接触がもたらす空間情報からなる．このような複雑な構造を持った情報に基づいた統計処理を進めることが，言語研究にも必要になりつつある．

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 国立国語研究所：〒 190–8561 東京都立川市緑町 10–2

能地論文は構文解析のうち，ここ 20 年間の教師なし構文解析や文法推定（Grammar Induction）の動向について紹介している．構文木を扱う枠組として，句構造文法・依存文法・組合せ範疇文法についての研究について，どのように文法規則を推定するかについて概説している．この試みは「文法がどのくらい生得的なのか」について，テキストデータから統計的に示唆を与えるものであり，今後の進展が期待される．

村脇論文は言語類型論（Typology）に対する統計的なアプローチについて概説している．近年，言語の類型論的特徴データが整備され共有されるようになった．これらのデータの特性を解説しながら，系統樹を復元する統計モデルについて紹介している．5 節では日本語の起源についても言及しており，クレオール形成に着目した言語接触に関する議論は特に興味深い．

岡崎論文は系列ラベリング技術に関するものである．岡崎氏は CRFSuite[2)] と呼ばれる条件付確率場（Conditional Random Fields）に基づく系列ラベリングツールを公開しており，系列に対するロジスティック回帰についての丁寧な解説を行っているほか，系列ラベリング研究の最新の動向についても言及している．

新井論文は心理言語学で用いられる眼球運動測定器から得られるセンシングデータをどのように統計処理するべきかを紹介している．眼球運動測定器から得られる情報がどのような性質をもち，どのようにモデリングするかを解説している．実験のデザインから解説しており，心理言語学の分野に興味がある研究者への良きチュートリアル資料になっている．

最後に荒牧論文は，ソーシャルメディアサービスを用いた言語研究について紹介している．ソーシャルメディアは非文法的でノイズの多い言語データであり，多数の発言者のデータを発言時刻と発言地点などの GPS 情報とともに得ることができ，これらを用いた新たなアプリケーションについて言及している．

こうした研究は，統計的機械学習の一部である統計的自然言語処理，および言語学自体への統計的手法の導入から生まれたものである．本特集は，その中でも特に工学より言語科学に近い，今後の発展が見込まれる内容を議論して論文の執筆を依頼した．一方で本特集は CRF や教師なし構文解析といった，ある程度確立された分野の，これまでになかった詳しい解説ともなっており，自然言語処理の技術に興味がある方にも有益な内容となっていると考えている．

前回の 2000 年の特集は，今回扱っている統計的自然言語処理とデータ科学の時代を予感させるものであった．今回の論文の内容をよく読むと，次の時代の研究の種子があちこちに埋まっていることがわかる．たとえば能地および村脇論文は本質的に，言語に存在する木構造や分岐構造を生成する確率過程をどう考えるか，という問題であるし，荒牧論文では暗黙に，感染症のモデルである SIR モデルが言語の場合にどう現れるかという問題を扱っている．どちらも統計学の分野でこれまで研究はあるものの，統計学の場合は問題を数理的に解きやすくするために大きく単純化されており，今回のような言語の場合にどう適用していくかはまだほとんど未開拓の荒野であるといってよい．

また新井論文では，眼球運動データと文の読み時間が分析されており，逆ガウス分布の当てはまりがよいことが示されている．逆ガウス分布を使う理由については引用されている Lo and Andrews (2015) の中でも弱い形で説明されているが，もし眼球運動を（バイアスのかかった）ブラウン運動とみなせば，その一定距離，すなわちここでは文末までの到達時間分布が逆ガウス分布になることは統計学においてはよく知られた事実であり，これは新井論文の観察とも符合する．したがって，視線の運動を直接，特殊なランダムウォークとしてモデル化することもできそうである．

岡崎論文は CRF（Conditional Random Fields，条件付確率場）の入門と詳細な解説であるが，論文の中で述べられているように，CRF はロジスティック回帰を時系列化したものといえ，現代の統計的自然言語処理全般においてきわめて重要な方法となっている．本特集ではふれなかった

が，現在ニューラルネットワークに基づく手法が自然言語処理を席巻しており，本特集編集時にも，Google 翻訳がニューラルネット化して大幅に精度が向上したことが話題となった．ニューラルネットがどのように動作しているのかはまだ解明されていないものの，このような中で，ニューラルネットで CRF の素性関数を学習する CNF（Conditional Neural Fields）(Peng et al., 2009)，画像の場合であるが CRF の学習をレイヤー化して再帰的ニューラルネットとして捉え，ニューラルネットとの統合を図る研究（Zheng et al., 2015）などの研究が発表されており，統計的に堅実な手法である CRF がどのように利用されていくのか，今後の動向が期待される．

こうした現在の統計的言語研究の先端を詳しく紹介する中で，その内容とともに，次の時代の言語研究の息吹を読みとっていただければと考えている．最後に，新しい試みが多い本特集の論文をお願いできる方を言語学または自然言語処理の中で探すのには労を要したが，貴重な査読者の方々にはお忙しい中，大変有益なコメントをいただくことができた．この場をお借りして感謝を申し上げたい．また，再度にわたる原稿の修正を行っていただいた執筆者の皆様，およびお世話になった編集室の方々にお礼を申し上げたい．

注．

1) http://www.ism.ac.jp/~daichi/workshop/2015-statling/
2) http://www.chokkan.org/software/crfsuite/

参 考 文 献

Lo, Steson and Andrews, Sally (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data, *Frontiers in Psychology*, **6**, 1171.

Peng, Jian, Bo, Liefeng and Xu, Jinbo (2009). Conditional neural fields, *NIPS 2009*, 1419–1427.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C. and Torr, P. H. (2015). Conditional random fields as recurrent neural networks, *ICCV 2015*, 1529–1537.