統計数理研究所 75 周年記念事業（2019）

統計的自然言語処理と統計学

Natural language processing and Statistics

数理・推論研究系 持橋 大地（Daichi Mochihashi）

1. 「静的な統計」から「動的な統計」へ

2011 年に統数研に准教授として着任後, 2016 年の『統計数理』64 巻 2 号において, 特集「統計的言語研究の現在」を企画・担当した. 言語に関しては『統計数理』では 2000 年の 48 巻 2 号で特集「“ことば” 新研究」が組まれており, 二者を比較すると, 16 年の間に大きな変化があったことに気づく. 2000 年の特集では文章のジャンルの分類や編集距離による類似和歌の発見, 多変量解析による文書の因子分析などが主な内容であったが, 2016 年では構文構造の教師なし学習, 言語変化への統計的アプローチ, CRF(条件付確率場) の詳細な解説など, 内容が大きく様変わりしている. 統計的な手法としても, 前者が多変量解析をベースとしているのに対し, 後者では系統樹の統計モデル, ポアソン過程, 条件付確率場 (ロジスティック回帰の隠れマルコフモデル化), 階層ベイズモデルのように最新の幅広い統計あるいは機械学習の手法が取り入れられるようになった.

手法以外にも, 研究の哲学ともよぶべきものが, 「静的な統計」から「動的な統計」へと変わったように感じられる. 前者では, 文や文章は確定した「モノ」であり, それをどう扱うかがテーマとなっていたが, 後者では言語自体も変化するものであり, 内部に多くの文脈依存性を持っていることがフォーカスされている. 例えば CRF は文の構文解析や, 各単語の品詞が互いに依存するマルコフ確率場のモデルであり, 特集に含まれている文の読み時間の動的な推定や Twitter のツイートのモデルも, 時間依存性や空間依存性を含んでいる.

現在, 「動的」とは主に, ある文や文章内の現象を動的にモデル化しているが, 上の言語系統樹の話にあるように, 世代を超えて言語自体が時間的・空間的に変わりうる様相を統計的に明らかにすることにも今後取り組みたく, 国立国語研究所や国立民族学博物館との共同研究を現在行っている.

2. 単語と分節化の理論

英語のように単語に分かれていない[1] 日本語や中国語, タイ語のような言語にとって, 文を「単語」に分けることは最も基礎となる重要な課題である. 従来はこのために, 人手で準備した大量の「正解」の単語列をもとに CRF などを学習することで単語分割や品詞推定がなされてきたが, 「正解」が真に正解であるという保証はなく, また日々無数に生まれる新語には対応できないという限界がある.

一の皇子は、右大臣の女御の御腹にて、寄せ重く、疑ひなきまうけの君と、世にもてかしづきこゆれど、この御にほひには並びたまふべくもあらざりければ、おほかたのやむごとなき御思ひにて、この君をば、私ものに思ほしかしづきたまふこと限りなし。はじめよりおしなべての上宮仕したまふべき際にはあらざりき。

図 1.『源氏物語』の教師なし形態素解析の結果の一例. 辞書や文法は一切用いていない.

[1] ラテン語や英語も, もともとは単語を分けて書かず一続きに書くのが普通であった.

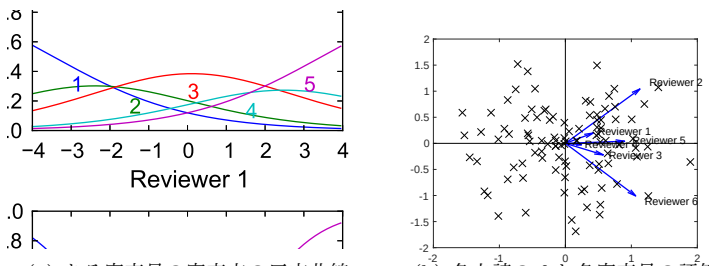
これに対し, ノンパラメトリックベイズ法による文字–単語の階層ベイズモデルを仮定し, 出力された生の文字列のみから「単語」を逆に推定する統計モデルを 2009 年頃に発表した (Mochihashi et al., 2009). これにより, 例えば「源氏物語」の文字列のみから単語を図 1 のように推定できる. その後の統数研での共同研究で, さらに品詞を同時に教師なし推定したり (Uchiumi et al., 2015), 教師データも利用した半教師あり学習も高精度で行えるようになった (Fujii et al., 2017).

統計的には, これはセミマルコフモデルの一種による分節化であると考えられる. したがって, 同様の統計モデルを言語だけでなく, 他の時系列データにも適用することができる. 言語において文字列にあたる出力をガウス過程からの波形とすれば, ロボティクスにおいてロボットの関節角の時系列を分節化して「動作」を取り出すことも可能になった (Nagano et al., 2018). 音声認識においても, 音声から単語を自動認識することが可能になるため, 共同研究を行っている.

3. 離散データと「科学の科学」

離散データを扱う統計的自然言語処理は, 他の多くの分野と繋がりを持っている. 2015 年から現在まで研究員を務めている日本学術振興会の学術情報分析センターでは, 科研費の申請に査読者が付した “5”,“2” などの審査点から, 項目反応理論により図 2 のように正規分布に従う各申請のスコアを統計的に算出する研究を行った. 多次元で行えば, 矢印で示した各審査員の「評価軸」も客観的なデータから数学的に明らかになり, 公正な審査に貢献する. また, 大規模な確率的潜在意味解析 (LDA) を用いて, 学振に登録されている研究者の専門性を計算し, 審査の際に適切な審査員を推薦する試みも行っている.

科学的な発見は論文, つまり言語の形で発表される. 論文の生産自体も離散的なイベントであり, 点過程として捉えることができる. また, そこには複雑な相互依存関係があると考えられる. 自然言語処理の背景を生かし, こうした「科学の科学」へも今後取り組みたいと考えている.



(a) ある審査員の審査点の反応曲線. (b) 各申請の θ と各審査員の評価軸.

図 2. 項目反応理論による潜在的な得点と評価軸の計算

参 考 文 献

Fujii, R., Domoto, R. and Mochihashi, D. (2017). Nonparametric Bayesian Semi-supervised Word Segmentation, *Transactions of ACL*, **5**, 179–189.

Mochihashi, D., Yamada, T. and Ueda, N. (2009). Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling, *Proceedings of ACL-IJCNLP 2009*, 100–108.

Nagano, M., Nakamura, T., Nagai, T., Mochihashi, D., Kobayashi, I. and Kaneko, M. (2018). Sequence Pattern Extraction by Segmenting Time Series Data Using GP-HSMM with Hierarchical Dirichlet Process, *IROS 2018*, 4067–4074.

Uchiumi, K., Tsukahara, H. and Mochihashi, D. (2015). Inducing Word and Part-of-speech with Pitman-Yor Hidden Semi-Markov Models, *ACL-IJCNLP 2015*, 1774–1782.