学会記事

「言語統計力学」＝言語学・自然言語処理・物理学

持橋 大地[†]・小木曽 智信[††]・高村 大也[†††]・小町 守[††††]

2022 年 3 月 9 日, 言語処理学会年次大会の少し前に, TKP 東京駅セントラルカンファレンスセンターと Zoom のハイブリッドで, 「現代語の意味の変化に対する計算的・統計力学的アプローチ」シンポジウムを開催しました.

このシンポジウムは同名の研究プロジェクトのまとめとして開催したもので, 184 人の参加登録と, 現地からの講演に加え, オンラインでは 100〜120 名の多数のご参加をいただきました. 関連した招待講演も含め, シンポジウムのページで現在も発表資料を公開しています.[1] 本稿ではこの研究プロジェクトの由来と, 研究を通して得られた知見について, 言語処理学会の皆様に紹介したいと思います.

1 プロジェクトの始まり

この共同研究は, 国立国語研究所の公募する「新領域創出型共同研究プロジェクト」に採択されたものですが, 当初は予算のサポートのある研究ではなく, まったくインフォーマルに始まったものでした. 代表者の持橋が 2018 年ごろ, 知り合いの SNS での会話をきっかけに, 言葉の意味の変化はそれぞれの言語使用者が周りに合わせて使い方を確率的に変えることで, 物理学でいうイジングモデル (MacKay 2003, 31 章) のようなことが起きており, 一種の相転移として意味の変化をとらえられるのではないか, と考えたのが始まりでした.

統数研は定員が少ないため, 特に研究テーマを振れる学生もいないことから, Twitter でこうした研究をしたい人はいませんかと募集したところ, すぐに何人かの方から応募がありました. そこで, 都立大の修士の竹中さんおよび, これまで関連のある研究 (Takamura et al. 2005) をしてきた東工大/産総研の高村氏と国語研の小木曽氏, 竹中さんの指導教員である小町氏などで研究を始めました. 何回か集まって検討するうち, 小木曽氏から国語研究所の「新領域創出型共同研究」に出してみてはどうかという提案があり, 意外にも高倍率でしたが, 幸いなことに採択していただけて, 3 年間の研究プロジェクトを続けることができました.

竹中氏は会社の仕事が忙しくなったこともあり, 学生は当時長岡技科大の学部 4 年生だった

[†] 統計数理研究所
[††] 国立国語研究所
[†††] 産業技術総合研究所
[††††] 東京都立大学

[1] https://www.ism.ac.jp/~daichi/workshop/2022-lchange/

相田さんと, その後で小町研に入学した修士 1 年の井上さんに引き継がれて現在に至っています. 現在も研究は続けていますので, 興味のある方は持橋までご連絡ください. また, この話題は統計力学と非常に関係が深いことから, 申請にあたり自然言語処理だけでなく, 統計力学を専門とする統数研の坂田綾香氏, および小山慎介氏にも加わっていただいたことが特徴です. こうして, 主に統計力学チームとデータ解析チームに分かれて研究を進め, ときおり全体ミーティングを行って進捗を共有していきました.

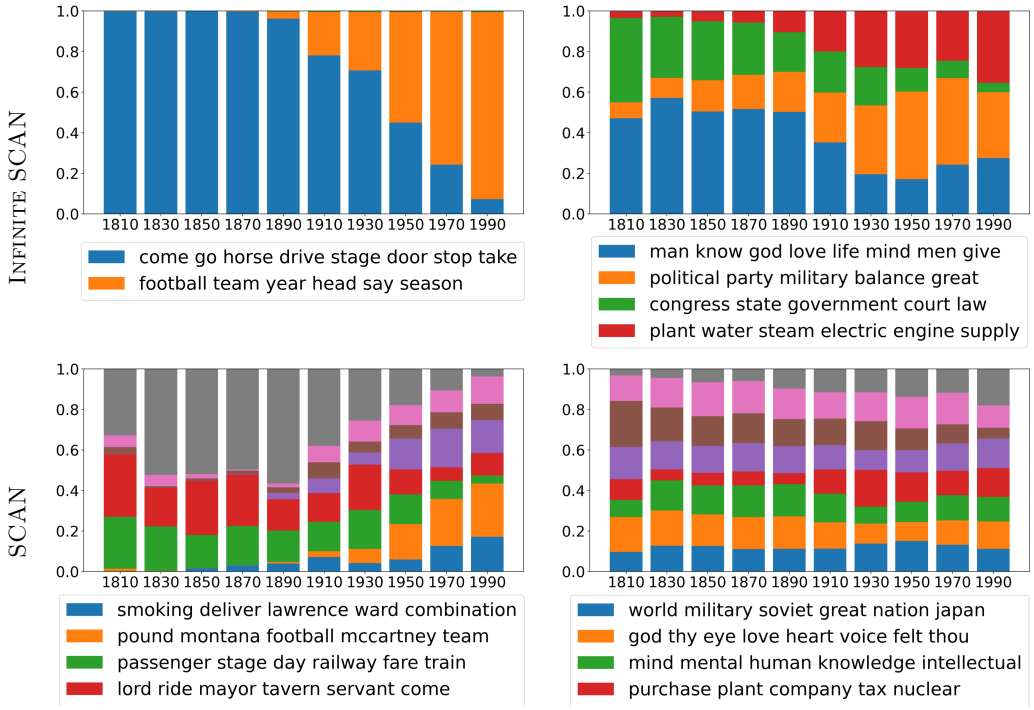
2　統計力学と自然言語処理

検討を始めて最初に気がついたのは, こうした, 言語変化を抽象的に扱う理論的な研究は, 言語学でも自然言語処理でもなく, 物理学で行われているということでした. Physical Review のような物理のジャーナルに興味深い研究が多数あり, 言語学で知られていた「言語変化は S 字カーブで進行する」という事実 (横山, 真田 2007) は, Bass モデルという形の微分方程式によって, すでに理論的な説明が得られていました (Ghanbarnejad et al. 2014). こうした研究を主に筆者が追い, 統計力学チームと共有する中で, 近似マスター方程式を使った解析や分岐過程による感染のモデル化の論文 (Keating et al. 2022, Physical Review E) にまでたどり着き, 言語変化の理論的な側面について, ほぼ最前線に近いところに達したと判断することができました.

ここで実験が行えるとよかったのですが, 残念なことに, これまでの物理学や言語学での言語変化のモデル化は, 綴りの変化や動詞の活用の変化といった, 客観的に変化がすぐ見えるデータを対象にしていました. しかし, 「意味」の変化は直接それ自体が目に見えるものではありません. そこで, まずは統計的分析により, 意味の変化をコーパスから統計的に取り出さなければなりません. そんなわけで, 後半から現在に至り, データ解析としての研究をプロジェクトでは主に進める結果になりました.

とはいえ, 言語変化について, 微分方程式を使ったモデル化でここまで数理モデルができていること, そしてそれが, 言語学者にも自然言語処理の研究者にもほとんど知られていないことは驚きで, かつ, サイエンスとしての可能性を感じた出来事でした. こうした研究には, 当然ながら深層学習のツールキットが使えるだけでは無理で, 物理的な知識がどうしても必要になります. 現在の自然言語処理にはそうした分析が相当に欠けており, 今後, 物理出身の方[2]を中心に, より NLP が深まるとよいのではと筆者は考えています.

[2] 筆者も学部 4 年生の 1 年間は物理系の研究室に属しており, その影響があるのかもしれません.



(a) "*coach*"　　(b) "*power*"

図 1　Infinite SCAN (上段) および SCAN (下段) による COHA コーパスからの意味変化の推定結果. 各トピックが 1 つの語義を表しており, 語義数を事前に固定する必要のある SCAN に比べて, 提案法ではそれぞれの単語の (変化する) 語義の数が自動的に学習されています. 図の下部は各トピックの上位語で, SCAN では一部のトピックのみを表示しています.

3　言語変化の研究

言語変化については多くの研究がありますが, 言語学の視点からみて最も欠けていると思えたのは, 「どんな単語が意味変化するのか」という網羅的な研究が少なく, 自然言語処理でも apple や gay といった代表的な単語の例だけで説明されていることでした. 単語ベクトルを使って網羅的に意味変化した単語を求めるには, 違った時期の単語ベクトルを対応づける必要があります. Hamilton(2016) はこの対応を線形変換で求めていたため, 制約が大きいと考えられました. そこで, 単語ベクトルは共起関係の PMI(自己相互情報量) 行列の行列分解と数学的に等価だという研究 (Levy and Goldberg 2014) を踏まえ, 二つの時期の共起行列の行列分解を同時に行うことで対応づけを不要にする方法を相田さんを中心に研究し, この研究は言語処理学会での発表 (相田他 2021) を経て, PACLIC 2021 に採択されました (Aida et al. 2021).

国語研の小木曽氏を通じて日本語学の研究集会でも発表を行う中で, 言語学の立場からは, 単に意味変化の大小だけでなく, 単語の意味がどう変わったのか, 何が変わったのかを知りたいと

いう要望が大きいことがわかりました. このためにはニューラル手法よりも, 異なる「語義」を潜在変数として明示的にモデル化するトピックモデルのような統計モデルが適しています. 動的なトピックモデルによって, 単語の変化する意味を追跡する SCAN というモデルが存在していましたが (Frermann and Lapata 2016), これは事前にトピック数, すなわち語義の数を指定する必要があるものでした. 語義の数はほとんどの場合, 事前にはわかりません. また指定する語義の数によって, 語義が重複したり, 本来は別の語義が同じトピックにまとめられてしまう可能性があります. そこで, ノンパラメトリックベイズ法によって, SCAN で変化する語義の数もデータから自動的に推定する統計モデル, Infinite SCAN を開発しました (図 1). SCAN はガウス分布のマルコフ確率場でモデル化されているため, この上で無限次元を表現できるロジスティック棒折り過程 (Ren et al. 2011) を利用しているのが特徴です. 複雑な計算が必要ですが, 担当した井上さんの頑張りによりモデルは完成し, 言語処理学会での発表 (井上他 2022) を経て, 現在は国際会議での発表を準備しています.

これらの研究には, 英語については COHA (Corpus of Historical American English) という 1820–2019 年の 4.7 億語のコーパスを用いています. 日本語にはこうした通時コーパスがこれまでなかったため, 研究には国立国語研究所で作成している近現代雑誌通時コーパス (近藤, 相田, 小木曽 2022) を使っています. これは, 1874–2013 年の間の『太陽』『中央公論』『文藝春秋』といった雑誌の記事を約 8 年おきに収録したもので, 約 4 千万語のコーパスとなっています. プロジェクトではこの国語研通時コーパスを用いた他の研究も行っており, 今後は日本語学や, 国語辞典の編纂を統計的にサポートするなどの応用が期待できます.

言語変化の研究は NLP の中でも注目度が高く, 競争が激しい領域ですが, その中で技術的に着実な成果を上げられたと考えています. 現在は単語ベクトルの空間の中で, どのような特徴をもつ単語が意味変化しやすいのかについて検討しています. また統計力学的分析のためには, 時系列に即した網羅的なデータ化が必要で, 潜在空間においてカルマンフィルタを考えることを次に予定しています.

4　まとめ

「言語統計力学」の国語研共同研究プロジェクトは, 自然言語処理, 言語学・国語学, 統計物理学の共同研究プロジェクトでした. 研究はまだ終わっていませんが, 言葉の変化の問題は言語学において普遍的な問題です. かつ, 意味的な変化は表立って見えるものではないため, 自然言語処理の技術が必ず必要になる研究であるといえるでしょう. 言語学者が言語処理学会に来なくなった, と言われて久しくなりますが, 言語学の若手には自然言語処理を学習しようとする動きが大きくあるようで, 自然言語処理の側でも, そうした方々と協力して次の時代の言語学を作っていく必要があるのではないかと筆者は考えています.

謝　辞

この研究は, 国立国語研究所の新領域創出型共同研究「現代語の意味の変化に対する計算的・統計力学的アプローチ」として行われました. メンバーの皆様, およびシンポジウムに参加いただいた皆様に感謝申し上げます.

参考文献

Aida, T., Komachi, M., Ogiso, T., Takamura, H., and Mochihashi, D. (2021). "A Comprehensive Analysis of PMI-based Models for Measuring Semantic Differences." In *PACLIC 2021*, pp. 21–31.

Frermann, L. and Lapata, M. (2016). "A Bayesian Model of Diachronic Meaning Change." *Transactions of the Association for Computational Linguistics*, **4**, p. 31–45.

Ghanbarnejad, F., Gerlach, M., Miotto, J. M., and Altmann, E. G. (2014). "Extracting information from S-curves of language change." *Journal of the Royal Society Interface*, **11**.

Keating, L. A., Gleeson, J. P., and O'Sullivan, D. J. P. (2022). "Multitype branching process method for modeling complex contagion on clustered networks." *Physical Review E*, **105** (3), p. 034306.

Levy, O. and Goldberg, Y. (2014). "Neural Word Embedding as Implicit Matrix Factorization." In *Advances in Neural Information Processing Systems 27*, pp. 2177–2185.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). "Logistic Stick-Breaking Process." *Journal of Machine Learning Research*, **12**, pp. 203–239.

Takamura, H., Inui, T., and Okumura, M. (2005). "Extracting Semantic Orientations of Words using Spin Model." In *Proc. of ACL 2005*, pp. 133–140.

井上誠一, 小町守, 小木曽智信, 高村大也, 持橋大地 (2022). ガウス確率場による単語の意味変化と語義数の同時推定. 言語処理学会第 28 回年次大会, pp. A7–1.

横山詔一, 真田治子 (2007). 多変量 S 字カーブによる言語変化の解析. 計量国語学, **26** (3), pp. 79–93.

近藤明日子, 相田太一, 小木曽智信 (2022). 近現代雑誌通時コーパスの語彙統計情報の公開. 言語処理学会第 28 回年次大会, pp. PT4–1.

相田太一, 小町守, 小木曽智信, 高村大也, 持橋大地 (2021). 通時的な単語の意味変化を捉える単語分散表現の同時学習. 言語処理学会第 27 回年次大会, pp. E4–3.

略歴

持橋 大地：1998 年東京大学教養学部基礎科学科第二卒業. 2005 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了. 博士 (理学). ATR 音声言語コミュニケーション研究所, NTT コミュニケーション科学基礎研究所各研究員を経て, 2011 年より情報・システム研究機構 統計数理研究所 准教授. 専門は自然言語処理および機械学習. 2015 年より, 日本学術振興会 学術情報分析センター分析研究員を兼務.