特　集「ポスト経験主義の言語処理」

統計的自然言語処理と機械学習・統計学の未来

Statistical Natural Language Processing and Statistics in the Future

持橋　大地　統計数理研究所モデリング研究系
Daichi Mochihashi　Department of Statistical Modeling, The Institute of Statistical Mathematics.
daichi@ism.ac.jp, http://www.ism.ac.jp/~daichi/

Keywords: computational linguistics, statistics, statistical machine learning.

1. はじめに

統計的自然言語処理あるいは機械学習と，古くからある「言語学」との関係を考え直す，というとき，よくある解答は，二つの分野は互いを尊重しつつ，独立に発展すべき，という優等生的な解答であろう．しかし，本当にそうだろうか．

著者は上の問題に対し，少なくとも理論言語学の面では，統計的自然言語処理は言語学の上位クラスであり，言語学を包摂すべき，と考えている．別の言い方をすると，統計的自然言語処理こそ現代の理論言語学であり，代数的な性質を調べる「言語学」はその一部分をなすのが自然である，といってもよい．この主張が強いと思われる方がもしいれば，その逆を考えてみればよい．記号的・代数的性質に単に連続値の重み付けをするだけで，以下で紹介するような最先端の統計的自然言語処理を網羅することは，全く不可能だと著者は考える．

もちろん，現実の言語データを調べる実証的研究は重要であり，文系の“言語学”として残っていくだろう．この状況は，長い歴史の中で理論物理学と実験物理学に分業されている物理学の場合と同様であり，統計的言語処理は，確率的ではあるが[*1]，「言語の物理」として，言葉の科学となっていくのではないだろうか．

以下，本稿では著者が上のように考える理由を，最新の統計的自然言語処理の広がりに触れつつ述べてみたい．

2. 「統計＝重み付け」からの脱却

PCFG（確率的文脈自由文法）に始まり，Stochastic HPSG [Brew 95], Stochastic CCG [Kwiatkowski 10] などがこれまで導入された立場からは，統計というものは「ルール」に「重み」を与えるものであり，それ以上でも以下でもないという見方が根強いように思える．しかし，現代の統計的自然言語処理は，それらをはるかに超えた地点にまで発展している．

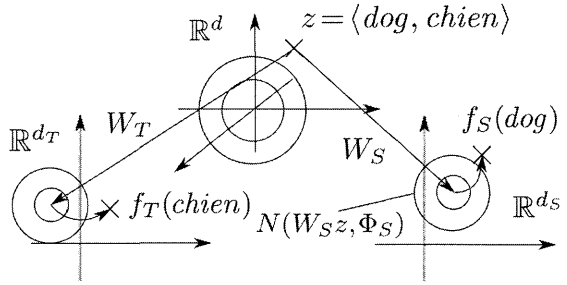

図1　潜在的CCAに基づく教師なし同義語の学習 [Haghighi 08]

例えば，Haghighiらによる潜在的な正準相関分析（CCA）を用いた教師なし対訳語対の学習 [Haghighi 08] では，二言語の無関係なコーパスから *dog*（英）-*chien*（仏）のような対訳単語対を得るために，この単語対が潜在空間において共通の座標 $z \in \mathbb{R}^d$ をもっていると仮定する．z の各言語への線形射影 $W_S z$, $W_T z$ を中心としたガウス分布 $N(W_S z, \Phi_S)$，$N(W_T z, \Phi_T)$ により，単語の共起情報や綴りの素性ベクトル f_S, f_T がそれぞれ生成されたと考えて，可能な対訳対を探索することで，二つの単言語コーパスのみから90%以上の高い精度で対訳対を得ることに成功した（図1）．

また，インド・ヨーロッパ祖語のような，言語の見えない系統樹を復元することは言語学の大きな夢であるが，最近の研究である [Bouchard-Côté 09] では，単語の

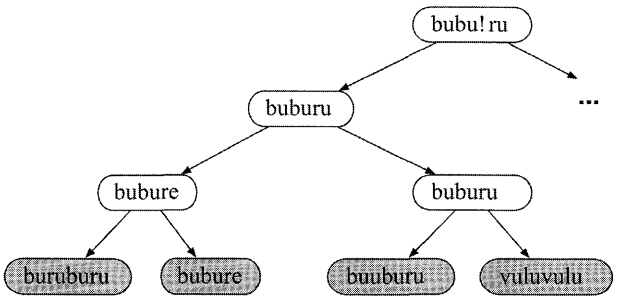

図2　[Bouchard-Côté 09] における綴りの系統樹の推定（オセアニア祖語）．
アミがけの各言語での観測文字列から，潜在的な祖先の文字列を次々とサンプリングして学習する

*1 言語ではミクロな振舞いも意思や環境の影響で確率的となるため，統計物理と完全に同一視できるわけではない．

スペルの系統樹を統計的に推定する（図2）．この研究では，各言語で同じ意味を表す単語集合から，それらが見えない原形からいかに進化したかを統計的に学習する．Coalescent tree [Kingman 82] などの確率論の成果が前提となっており，「いつ文字列の変異が起こったのか」が指数分布に従うとしたうえで，最終的に得られた文字列から，その潜在的な親と祖先をサンプリングするモデルになっている（これには高度なMCMC法を要する）．

これらはすべて，「規則の重み付け」という観点からは扱えないのは明らかであろう．ほかにも，統計的自然言語処理には強化学習やカーネル法，無限モデルなどが機械学習の一部として，最近次々と導入されている．現代の統計的言語処理は，線形空間や確率過程といった数学の枠組みの上に，次々と発展を続けている．

なお，古典的な言語学で扱えない「重み付け」や“gradedness”自体についても，それが単に語用論の問題として片付けることのできない，いかに不可欠なものであるかについて，ともに言語学出身のPereira [Pereira 00] やAbney [Abney 96] が詳細に解説しているので，興味のある方はぜひ参照されたい．

3. 「深い理解」とは？

もう一つのよく言われる言語学必要論は，言語の「深い理解」のためには言語学者の与えた文法やアノテーションが不可欠，とするものである．しかし，これは慎重に検討する必要がある．

確かに，生の言語データのみから得られる情報には限界があるのは事実である．テキストは裏にある膨大な体験，事実をもとに生成された，言ってみれば現実界の「写像」にすぎないからである（ヴィトゲンシュタイン的な意味で）．つまり，テキストからその「原像」を得るためには補助情報，統計的には共変量（covariate）の存在が非常に重要となる*2．このとき，その補助情報が，言語学者の主観によるラベルである必然性はない．言語学者が「正しい」保証はどこにもないからである．むしろ，自然科学の立場からは，意図的に付けたものではない，自然に得られる共変量を重視すべきであろう．

例えば，単語が正の意味か負の意味かを知るのに，単語自体を手でタグ付けする必要はなく，Amazonのレビューで星の数の多い文章には正の意味の単語が統計的に多く含まれている，ことを前提とした統計モデルにより，共変量である評価値から単語のもつ正負を推定することができる（近い話として，[Yoshida 11]）．また，何が「単語」であるかを知るのに言語学者の定義に頼る必要はなく，計算機に入力する文字のタイミングや，会話の録音データを用いればよいし，“お茶漬でも”という文の意図を知るには，前後の状況を使うことができるだろう．こうした共変量の利用可能性は，現在の新聞記事に基づいた自然言語処理のパラダイムをいったん超えれば，ほぼ無尽蔵といってよいほどに残されている．もちろん，その統計モデル化は簡単ではないが，方法を開発することができれば，極めて自然で，有効な解釈がもたらされるはずである．

3·1 意味の問題

同様なものに「意味」の問題がある．文や文章の表層を超えた意味を理解するためには，「意味的アノテーション」が不可欠，とするものであるが，それが唯一の方法だとは限らず，アノテーションなしで同様のことができれば理想的なのは明らかであろう．実際，[Liang 11] では，従来手で与えていたλ計算の論理式に替え，DCS（Dependency-based Compositional Semantics）と呼ばれる依存構造を潜在構造として入力文 $\boldsymbol{w}$ と解答 y のペア $\langle \boldsymbol{w}, y \rangle$ のみから教師なし学習をすることによって，人手による高価なアノテーションなしに，質問応答での大幅な高性能化を達成した（図3）．彼らのモデルでは，DCS z を潜在変数とみなして，解答 y を与える z を可能な指数的な組合せ $Z_L(\boldsymbol{w})$ の中から探索し，下の目的関数を最大化するようにデータ D から学習する．

$$O(\theta)=\sum_{\langle \boldsymbol{w}, y\rangle \in D} \log p_\theta([\![z]\!]=y \mid \boldsymbol{w}, z\in Z_L(\boldsymbol{w}))-\lambda\|\theta\|^2 \tag{1}$$

$[\![z]\!]$ は z を評価する，という意味である．

一般に，「意味」とは状況によって変わり，最終目標はそれにより何らかの処理 y を可能にすることなのであるから，こうした意味的処理は，隠れた「意味」を積分消去する予測

$$p(y|\boldsymbol{w})=\int p(y,\theta \mid \boldsymbol{w})\,d\theta=\int p(y \mid \boldsymbol{w},\theta)p(\theta|\boldsymbol{w})\,d\theta \tag{2}$$

を行っていることになる．このとき，第1項の尤度 $p(y \mid \boldsymbol{w},\theta)$ および第二項の事前分布 $p(\theta \mid \boldsymbol{w})$ の設計が，客観的に制御可能な，意味の統計モデルとなるだろう．

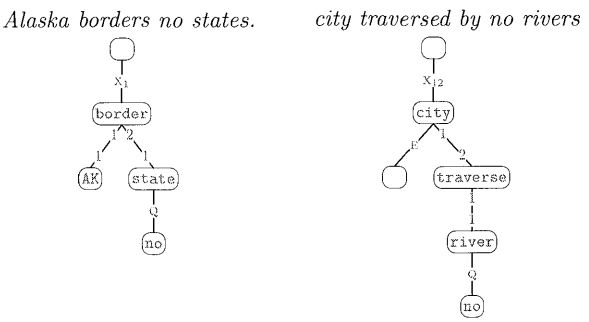

図3 [Liang 11] で使われているDCSの例

*2 ここで慎重に，「不可欠」とは言っていないことに注意されたい．程度の差はあれ，さまざまな方向から映した“影絵”が大量にあれば，そこから元の形を推定することは原理的に可能である．

4. 半教師あり学習と統計モデルの設計

上では主に教師なし学習，あるいは共変量を用いた学習について述べてきたが，従来の言語学や自然言語処理の蓄積はもちろん，非常に有用である．それらの教師情報を生かしつつ，大量の生データからの統計からの融合を図るものとして半教師あり学習があり，現在その目的関数は，$\langle Y_s, X_s \rangle$ を教師ありデータ，X_u を教師なしデータとして，

$$p(Y_s, X_s, X_u) \propto p(Y_s \mid X_s)^{\mu} \cdot p(X_u \mid Y_s, X_s)^{\lambda} \quad (3)$$

の形の積モデル（Products of Experts）[Suzuki 08]，または

$$-\log p_\theta(Y_s \mid X_s) + \lambda U(\theta) + \gamma H(p_\theta(Y \mid X_u)) \quad (4)$$

のような X_u を用いた正則化 [Jiao 06] などにより実現されている．ここで，H はエントロピー関数，U は L_2 などの正則化項を表す．

しかし，どちらもまだ十分に成熟した方法とはいえず，「教師データ」が必ずしも正解とはいえない場合も含め，教師データの取入れ方には課題が多く残されている*3．教師データのもう一つの使用法は，モデル設計に用いる，というものであろう．これは式（2）の設計に相当する．例えば，文の受身化や，“move α”はどう確率化すればよいのだろうか．CFG を超える文法の確率化と学習法の開発も含め，言語学的理論の確率化はまだほとんど手のついていない，未開の荒野であるといってよい．その荒野を歩くとき，従来のように単に和を1にするといったレベルの‘確率モデル’ではなく，数学の成果を背景に，確率過程に基づいた精密な定式化が必要になるだろう．

5. 広がる統計モデルと自然言語

上で触れた以外にも，現代の統計的自然言語処理は従来の「言語学」で扱えない，さまざまな発展を見せている．例えば，twitter の発言の確率モデル [Ritter 10] や，照応解析のための確率過程 [Rao 09] などは良い例であるし，今まで実証的に扱われていた地理言語学も，空間統計学 [Cressie 93] との統合により統計化されていくだろう（最近の研究に [Eisenstein 10] などがある）．

しかし，と古い言語学者は言うかもしれない．意味役割やシンボルグラウンディングなどの，文の「深い理解」はどうした？ と．

ところが，近年の統計のオンライン学習の発展を鑑みると，本稿の2章で述べた意味で，外界の画像や音，センサ情報を共変量として，言葉のシンボルグラウンディングや役割の統計学習が可能な公算は大きい（例えば，[Iwahashi 10, Kollar 10]）．そのとき，言語学者の主観と違った意味で，環境から計算的に，言葉の「深い理解」が可能になると思われる．

言語はそもそも，コミュニケーションのための道具として発達したものである．旧来の言語学は曖昧性のない，言語の代数的・離散的性質を求めてきたが，自然現象としてみるとき，それは言語の目的とは違っている可能性が高い．曖昧性の存在が，かえってコミュニケーション効率を高めているという最近の指摘 [Piantadosi 12] もあり，言語が本来もつ目的関数に着目するのが，すなわち統計的自然言語処理だともいうことができるのではないだろうか．

本章最後に，ノーバート・ウィーナーの下の言葉 [Wiener 79] を引用しておきたい．

> 文法はもはや第一に規範的なものではない．それは事実的なものになったのである．問題は，どんな符号をわれわれは使用すべきかということではなく，どの符号をわれわれが使用するかである．言語を詳細に研究してゆくと，規範的な問題が確かにでてきて，それらは非常に微妙なものであることは，全く真実である．しかし，それらは通信（コミュニケーション）の問題に関しては最終の細かい仕上げに当たるものであって，その最も基礎的な段階に当たるものではない．
> —ウィーナー：人間機械論 “IV 言語の仕組みと歴史”．

6. 展望とまとめ

統計的自然言語処理が Web の発達などに伴って発展を始めたのは1990年代後半であり，まだその歴史はたった十数年しかない．この間にも，自然言語の統計的手法は大きな進歩を見せてきた．例えば，90年代に導入された最大エントロピー法（ME）は最初，学習に Iterative scaling などを必要とする非常に重い計算モデルであったが，その後 ME は対数線形モデルの最尤推定と等価なことが知られるようになり，計算も L-BFGS のような最適化法の導入 [Sha 03] やその後のオンライン学習法の進歩により劇的に高速化され，確率モデルのコンポーネントに使われるようにもなった [Berg-Kirkpatrick 10]．

また，かつて大きな問題であった，HMM や PCFG などの次元数決定問題は，[Neal 03] などに始まった Infinite Models により，理論的にはほぼ解決されたといえ，計算技法の面でも基本的な EM アルゴリズムから，EM で解けない階層モデルに対する変分ベイズ法，EP 法，MCMC [Bishop 08] など，次々に高度化されて，学習可能なモデルの範囲は年々広がりを見せている．

---

*3 単純に教師データをベイズ的な事前知識として用いる方法は，一定の性能はもつものの，教師なしデータが大量にある場合は尤度関数が事前分布を圧倒してしまい，ほとんど教師なし学習で結果が決まってしまう，ということに注意されたい．

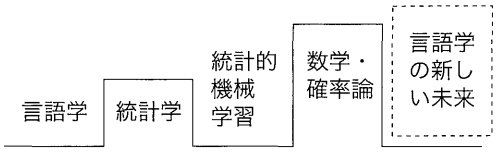

図4 言語学の未来と数学・確率論・統計学

さらに，かつて言語にとって不可侵の領域であった‘意味’にも，1999年の確率的潜在意味解析（PLSI）[Hofmann 99]，そのベイズ化であるLDAと後続モデルにより，いかに多くのことが可能になったかは記憶に新しい．

これらの高度な現代的な展開は，「言語学者」からは想像もつかないものであろう．単に人の与えた規則に重み付けをするだけの素朴な統計モデルをもって，「統計には限界がある」などと考えるのは，統計科学への無理解によっており，全くの早計といえるのではなかろうか．

言語学への統計学の導入により，統計的機械学習の一部にもなった統計的自然言語処理，あるいは計算言語学は，さらに深い数学や確率論との結び付きにより，本稿で紹介した課題を克服し，新たな楽園に入ることができると著者は考えている（図4）．その壁は高いが，高い壁もいつかは乗り越えられるはずである．

本稿によって少しでも，そうした意味での自然言語の統計モデルの可能性の広がりを感じていただけたらと願っている．

◇ 参 考 文 献 ◇

[Abney96] Abney, S.: *Statistical Methods and Linguistics*, pp. 1-26, MIT Press (1996), http://www.vinartus.net/spa/95c.pdf

[Berg-Kirkpatrick 10] Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J. and Klein, D.: Painless unsupervised learning with features, *NAACL-HLT2010*, pp. 582-590 (2010)

[Bishop 08] Bishop, C. M., 元田 浩，栗田多喜夫，樋口知之，松本裕治，村田 昇 監訳：パターン認識と機械学習：ベイズ理論による統計的予測（上）（下）(*Pattern Recognition and Machine Learning*), Springer (2007, 2008)

[Bouchard-Côté 09] Bouchard-Côté, A., Griffiths, T. L. and Klein, D.: Improved reconstruction of protolanguage word forms, *HLT-NAACL 2009*, pp. 65-73 (2009)

[Brew 95] Brew, C.: Stochastic HPSG, *EACL 1995*, pp. 83-89 (1995)

[Cressie 93] Cressie, N.: *Statistics for Spatial Data*, Wiley Series in Probability and Statistics (1993)

[Eisenstein 10] Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P.: A latent variable model for geographic lexical variation, *EMNLP 2010*, pp. 1277-1287 (2010)

[Haghighi 08] Haghighi, A., Liang, P., Berg-Kirkpatrick, T. and Klein, D.: Learning Bilingual Lexicons from Monolingual Corpora, *ACL 2008*, pp. 771-779 (2008)

[Hofmann 99] Hofmann, T.: Probabilistic latent semantic indexing, *Proc. SIGIR '99*, pp. 50-57 (1999)

[Iwahashi 10] Iwahashi, N., Sugiura, K.,Taguchi, R., Nagai,T. and Taniguchi,T.: Robots that learn to communicate: A developmental approach to personally and physically situated human-robot conversations, *Proc AAAI Fall Symp. on Dialog with Robots*, pp. 38-43 (2010)

[Jiao 06] Jiao, F., Wang, S., Lee, C.-H., Greiner, R. and Schuurmans, D.: Semi-supervised conditional random fields for improved sequence segmentation and labeling, *COLING/ACL 2006*, pp. 209-216 (2006)

[Kingman 82] Kingman, J.: The coalescent, *Stochastic Processes and their Applications*,Vol. 13, No. 3, pp. 235-248 (1982)

[Kollar10] Kollar, T., Tellex, S., Roy, D. and Roy, N.:Toward understanding natural language directions, *HRI 2010*, pp. 259-266 (2010)

[Kwiatkowski 10] Kwiatkowski, T., Zettlemoyer, L., Goldwater, S. and Steedman, M.: Inducing probabilistic CCG grammars from logical form with higher-order unification, *EMNLP 2010*, pp. 1223-1233 (2010)

[Liang 11] Liang, P., Jordan, M. and Klein, D.: Learning dependency-based compositional semantics, *ACL-HLT 2011*, pp. 590-599 (2011)

[Neal 03] Neal, R.: Introduction to infinite models, *NIPS 2003 Workshop of Nonparametric Bayesian Methods and Infinite Models* (2003)

[Pereira 00] Pereira, F.: Formal grammar and information theory: Together again?, *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, Vol. 358, No. 1769, pp. 1239-1253 (2000), http://www.cis.upenn.edu/~pereira/papers/rsoc.pdf

[Piantadosi 12] Piantadosi, S., Tily, H. and Gibson, E.: The communicative function of ambiguity in language, *Cognition*, Vol. 122, No. 3, pp. 280-291 (2012)

[Rao 09] Rao, V. and Teh, Y.: Spatial normalized gamma processes, *NIPS 2009* (2009)

[Ritter 10] Ritter, A., Cherry, C. and Dolan, B.: Unsupervised modeling of twitter conversations, *NAACL HLT 2010*, pp. 172-180 (2010)

[Sha 03] Sha, F. and Pereira, F.: Shallow parsing with conditional random fields, *HLT-NAACL 2003*, pp. 134-141 (2003)

[Suzuki 08] Suzuki, J. and Isozaki, H.: Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled Data, *Proc. ACL:HLT 2008*, pp. 665-673 (2008)

[Wiener 79] Wiener, N. 著，鎮目 恭夫，池原止戈夫 訳：人間機械論—人間の人間的な利用，みすず書房（1979）（原著 :Wiener, N. : *Cybernetics: or the Control and Communication in the Animal and the Machine*, MIT Press (1965)）

[Yoshida 11] Yoshida, Y., Hirao, T., Iwata, T., Nagata, M. and Matsumoto, Y.: Transfer learning for multiple-domain sentiment analysis: Identifying domain dependent/independent word polarity, *AAAI 2011* (2011)

2012年3月11日 受理

著 者 紹 介

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研究員を経て，2011年より統計数理研究所モデリング研究系准教授．統計的自然言語処理および統計的機械学習に興味をもつ．情報処理学会，電子情報通信学会，日本統計学会，ACL各会員．