4章 統計的自然言語処理と機械学習

持橋大地[†]

キーワード：自然言語処理，統計的機械学習，結合モデル，意味解析

1. まえがき

メディアデータとして，映像，すなわち静止画や動画の量は爆発的な成長を遂げている．離散的記号列である言葉は，一見それらと趣を大きく異にするようにみえるが，こうした映像は言葉と独立な存在というわけではなく，むしろ，言葉によって濃密に裏打ちされているといっていよい．多くの映像表現にはその前に言葉によって書かれたシナリオ，台本が存在するし，結果としての映像表現自体も，ニコニコ動画におけるコメント群や電子掲示板での実況中継に代表されるように，言語によってその内容を補完される運命にある．岩波ホールで最近公開された，沈黙の誓いを守り続けるフランスの修道院を描いたドキュメンタリー『大いなる沈黙へ』にあってすら，筆者の予想に反して言葉は存在し，訥々と語られるナレーションと，休日のみ許される修道士たちの自然な会話は，人間は言語なしには生きていけないことを象徴するかのようであった．

言葉を計算機上で扱う分野を自然言語処理，特に最近ではデータからの学習や曖昧性の解決のために不可欠となる統計を用いた言語処理を，統計的自然言語処理という．現在，自然言語処理ではまったく利用されていないが，しばらく前からテレビ番組の台詞にはほぼ字幕がつくようになり，音声に含まれる言葉も含めて考えれば，データとして利用可能な，言葉と映像の間の膨大な可能性をわれわれの前に提示してくれているように思われる．

そこで本解説では，そのために必要となる統計的自然言語処理の基本的な技術，および最近になって研究が進展しつつある，画像や動画と言語との連携モデルについて見ていくことにしよう．

2. 形態素解析，NER

東北海道でみられる鳥

のような字幕の日本語を見たとき，計算機は言葉を知らないから，まずこれを

[†] 統計数理研究所 数理・推論研究系

"Machine Learning in Statistical Natural Language Processing" by Daichi Mochihashi (The Institute of Statistical Mathematics, Tokyo)

図1 文字列「東横浜です」の形態素解析ラティス
(「東-横浜-です-ね」が正解のパスである)

東／北海道／で／み／られる／鳥

のように単語に分割し，必要に応じて名詞，動詞，形容詞などの品詞を付与する処理が最初に必要になる．これを形態素解析という．形態素解析は欧米語の場合には必要がないが，日本語やタイ語，十億を超える話者をもつ中国語などにおいては必要不可欠な処理である．単純な最長一致では，上の例は「東北／海道／でみ／られる／鳥」のように誤った解析になってしまう可能性があるため，言語情報に基づいた統計モデルが必要になる．

統計的には，これは L 文字の入力文 $\mathbf{s} = c_1 c_2 \cdots c_L$ (c は文字) に対し，それを T 個の単語に分けた単語列 $\mathbf{w} = w_1 w_2 \cdots w_T$ と，対応する品詞列 $\mathbf{t} = t_1 t_2 \cdots t_T$ を推定する問題と考えることができる．学習データとして与えられた文字列の集合 $S = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$ に対して，その単語分割 $\mathbf{w}$ や品詞列 $\mathbf{t}$ の正解データがわかっているとき (教師あり学習とよばれる)，これは系列データに対する分類問題であり，図1のように $\mathbf{s}$ に対して可能なすべての単語列を表すラティスに対して，太線で表した正解のパスに高い確率を，不正解のパスに低い確率を与えるように状態遷移確率を学習する機械学習の問題になる．具体的には，正解データの確率

$$p(\mathbf{w}, \mathbf{t} | \mathbf{s}) \quad (1)$$

を最大にするような確率モデルを学習する．

確率モデルはHMMを用いることもできるが，現在では対数線形モデル (ロジスティック回帰) によって

$$p(\mathbf{w}, \mathbf{t} | \mathbf{s}) = \frac{\exp(\Lambda \cdot F(\mathbf{w}, \mathbf{t} | \mathbf{s}))}{Z} \quad (2)$$

で計算する方法が、文字列や品詞列のさまざまな情報を文字列の特徴ベクトル F に含められて高性能なために、標準的に使われる方法となっている。ここで $F(\mathbf{w}, \mathbf{t} | \mathbf{s})$ はパスの特徴ベクトル、 $\Lambda = \{\lambda_1, \dots, \lambda_K\}$ は対応する回帰係数で、 Z はパス全体の確率を1にする正規化項である。実際には、式 (2) を系列に対する積の形に分解する。

$$p(\mathbf{w}, \mathbf{t} | \mathbf{s}) = \frac{\exp\left(\sum_{i=1}^T \sum_k \lambda_k f_k(\langle w_i, w_{i-1} \rangle, \langle t_i, t_{i-1} \rangle | \mathbf{s})\right)}{Z} \quad (3)$$

ここで $f_k(\langle w_i, w_{i-1} \rangle, \langle t_i, t_{i-1} \rangle | \mathbf{s})$ は $\mathbf{s}$ において単語・品詞の遷移が $\langle w_i, w_{i-1} \rangle \rightarrow \langle t_i, t_{i-1} \rangle$ であった場合に1、それ以外では0を返す特徴関数で、 k はその特徴のインデックスである。式 (3) の形に分解することで、学習にはHMMと同じ動的計画法を用いることができ、図1において各パスを通る確率の期待値から、各 λ_k に対する微分を計算して Λ を最適化することができる。これをConditional Random Fields (CRF)¹⁾ という。標準的なツールとして奈良先端大 (NAIST) で開発されたHMMベースのモデルである茶筌が有名であったが、現在は同大で開発されたCRFベースのMeCab²⁾ が広く使われている。

こうして、文字列「ノーベル文学賞」を解析して「ノーベル／文学／賞」が得られたとしよう。しかし、「ノーベル文学賞」は実際には一つの言葉であり、文が最初から単語に分かれている英語のような場合も含め、固有名詞とそのカテゴリをノーベル文学賞 (固有名詞) のように指示する処理をする必要が多く、これをNER (Named Entity Recognition) という。「New York City Philharmonic」のように複数の解析候補がありうる場合も多く、これにも機械学習が必要である。通常、Named Entityの最初の単語にB、内部にI、それ以外の単語にOを割当てて「BIOタグ」を学習データに付与し、上のCRFやSVMなどを用いてB、I、Oを予測する分類器を学習する。上の例では、ORGを組織名としてNew (B-ORG) York (I-ORG) City (I-ORG) Philharmonic (I-ORG) のようにタグが付与されて、学習が行われる。

3. 構文解析

3.1 教師あり構文解析

形態素解析およびNERを経て、「彼／が／望遠鏡／を／買った」のように単語列が得られても、それでは充分でない場合が多い。通常の時系列データと異なり、言語は複雑な構文構造を持っているからである。上の文に対して、図2のような構文解析がなされて初めて、われわれはこの文の主語が「彼」であり、目的語「望遠鏡」を「買った」ということがわかる^{*1}。

いま、図3のような確率的文脈自由文法 (PCFG) がわかっ

図2 PCFGによる構文解析の例

文法規則	確率
$S \rightarrow NP VP$	0.7
$S \rightarrow N NP$	0.3
$NP \rightarrow N P$	0.9
$NP \rightarrow P VP$	0.1
$VP \rightarrow 咲く$	1.0
$N \rightarrow 向日葵$	1.0
$P \rightarrow が$	0.6
$P \rightarrow に$	0.4

図3 確率的文脈自由文法の例

(S : 文, NP : 名詞句, VP : 動詞句, N : 名詞, P : 助詞)

ているものとしよう。学習データの各文がPCFGに従って構文解析されていたとすると、例えば、 $S \rightarrow NP VP$ となる確率は、 $n()$ を学習データでの頻度を返す関数として

$$p(S \rightarrow NP VP) = \frac{n(S \rightarrow NP VP)}{n(S)} \quad (4)$$

のように計算することができる。つまり、 $S \rightarrow NP VP$ や $S \rightarrow N NP$ といったSの展開可能性の中で、 $S \rightarrow NP VP$ と展開される相対頻度を求めているわけである。

上のようにして確率が求まれば、例えば「向日葵が咲く」が $(S (NP (N 向日葵) (P が)) (VP 咲く))$ となる確率は $p(S \rightarrow NP VP) \cdot p(NP \rightarrow N P) \cdot p(N \rightarrow 向日葵) \cdot p(P \rightarrow が) \cdot p(VP \rightarrow 咲く) = 0.42$ 、 $(S (N 向日葵) (NP (P が) (VP 咲く)))$ となる確率は同様に0.018、などと求まり、前者の方が確率のずっと高い構文解析であることがわかる。

ただし、上の計算は文法および学習データの構文解析結果が既知の場合のものである。英語には100万文程度のこうしたコーパス (言語処理用の学習データ) があるが、他の言語には小規模なもの、あるいはまったく存在しないことも多く、文法自体も英語の場合も含め、自明ではない。さらに、こうしたデータは多くの場合新聞記事を対象にしており、一般のメディアにおける口語や、新語を含んだ実際の言語現象は扱えない場合が多いことも問題となっている。

3.2 教師なし/半教師あり構文解析

これに対して、式 (4) のような確率を、構文解析自体と同時に推定しようとする方法の研究も進められている。既知の構文解析のような教師データを仮定しないため、こ

*1 こうした意味役割を直接求めるSemantic Parsingという方法もあり、研究が進められている。

$$p(S \rightarrow VP NP) = \frac{n(S \rightarrow VP NP) - td}{n(S) + \alpha} + \frac{\alpha + Td}{n(S) + \alpha} p_0(S \rightarrow VP NP) \quad (5)$$

係り受け解析：なお、図4のように、構文解析に係り受け関係として表すこともできる。これを係り受け解析という。係り受け解析木は適切な変換によってPCFGとして扱うことができるが、PCFGと異なり、①係り関係のみで定義されているため、語順に関して自由度が高く、②文の各単語について、その係り先の整数を割当てただけの組合せ問題である、という特徴がある。こうした特徴は学習について有利であることから、多数の解析アルゴリズムが存在して研究が進んでいる。

計算機により可能な意味処理の一つは、純粹に統計的なものである。例えば、画像にその中にあるものを指示する単語“tiger”, “grass”, “airplane”, “sky”などが付与されたデータセットを対象としてみよう。こうしたデータは、写真つき記事や字幕つき映像からも、ノイズは多いものの、容易かつ大量に得られると考えられる^{*2}。

彼 が 望遠鏡 を 買った

1	2	3	4	5	: 単語番号
2	5	4	5	0	: 係り先

これを式で表すと、 α, β, η をモデルのパラメータとして、単語集合 W および画像特徴量集合 F は z_a の集合 Z および θ を周辺化して

図7 動画の各領域に対するFHMMの学習
(口絵カラー参照)

図8 HMMとして表した言葉“towards”の意味

$$\begin{aligned} p(W, F | \alpha, \beta, \eta) &= \int \sum_Z p(W, F, Z, \theta | \alpha, \beta, \eta) d\theta \\ &= \int p(\theta | \alpha) \sum_Z p(W | Z) p(F | Z) p(Z | \theta) d\theta \\ &= \int p(\theta | \alpha) \prod_n \sum_{z_n} p(w_n | z_n) p(f_n | z_n) p(z_n | \theta) d\theta \end{aligned} \quad (6)$$

の確率によって生成されたとする。ディリクレ分布のパラメータ α および、話題 z 別の単語分布 $\beta = \{p(w | z) | z = 1, \dots, K\}$ と画像特徴量分布 $\eta = \{p(f | z) | z = 1, \dots, K\}$ がパラメータである。式 (6) は複雑なため解析解は得られないが、変分ベイズ法⁸⁾、マルコフ連鎖モンテカルロ法⁹⁾ などにより推定が可能である。上のモデルを学習することは、各単語がどの画像領域から生成されたかを確率的に推定することを意味するため、図5のような、画像と言葉との確率的対応づけが得られる。

上の手法は動画には対応していないが、やや違ったアプローチとして最近¹⁰⁾では、動画のトラッキングされた各領域について、その特徴量を言葉と結びつけることで、言葉の意味を図8のようにHMMとして学習する方法が提案されている。

動画に対応する文の、複数の言葉が同じ動画領域に関係しているため、これは複数のHMMの積が出力を共有するFactorial HMM¹¹⁾であり、どの領域がどの言葉に対応するのかも未知の潜在変数として学習が行われる。これにより

図9 “What is the most populous city in California?”の意味を表すDCS木 (口絵カラー参照)

図8のように、静的な言葉だけでなく、“towards”のように心理的に動作を伴った言葉の意味を、速度分布と目的物との距離分布の状態遷移として表すことが可能になり、注目されている。

5. 意味解析 (II)

さらに、最近文の持つ論理的構造に焦点を当てた、深い意味解析も現実的に可能な段階に入りつつある。例えば、“What is the sum of the population of ten largest cities in US?”という文に対して回答することや、「オスマン帝国では親衛隊イエニチェリが大きな役割を果たした。」という文について「トルコには軍隊が存在した」が真であることを示すのは、表層的な推定では不可能である。

従来これには、文献12)のように学習データの各文にその意味を表す λ -式 (論理式) を付与するという、きわめて作成の困難なデータが必要であった。しかし最近、文献13)により、意味構造自体を隠れ変数として学習し、 λ -式のような教師データを必要とすることなく、きわめて高精度な推論を行うことができる方法が提案されている。

この方法では、データとして x : “What is the capital of California?” $\rightarrow y$: “Sacramento” のような質問文 x と回答 y のペアの集合のみを与え、見えない意味構造 z を介して、 x が与えられたときの y の確率 $p(y | x)$ を最大化するように学習を行う。

$$\arg \max_{\theta} p(y | x) = \sum_z \underbrace{p(y | z)}_{\text{論理推論}} \underbrace{p(z | x, \theta)}_{\text{確率モデル}} \quad (7)$$

z は図9のような木構造で、これをDCS (Dependency-based Compositional Semantics) 木という。 z が与えられれば、 $p(y | z)$ はDCS上に定義された公理と推論規則から導出されるため、入力文 x が与えられたときのDCS木 z の学習 $p(z | x, \theta)$ が統計的な問題である。これにはロジスティック回帰を用い、 x から z が導出される特徴関数のベクトル $F(z, x) = (f_1(z, x), \dots, f_K(z, x))$ を使って

$$p(z | x, \theta) \propto \exp(\theta \cdot F(z, x)) \quad (8)$$

を学習データについて最大化するように、DCS木 z の指数的な可能性について動的計画法を用いて周辺化することで

	自然言語表現	外延の抽象表現
複合名詞	<i>pet fish</i>	$\text{pet} \cap \text{fish}$
修飾	<i>nice day</i>	$\text{day} \cap (W_{\text{ARG}} \times \text{nice}_{\text{MOD}})$
時間	<i>boys study at night</i>	$\text{study} \cap (\text{boy}_{\text{SBJ}} \times \text{night}_{\text{TIME}})$
関係詞節	<i>books that students read</i>	$\text{book} \cap \pi_{\text{OBJ}}(\text{read} \cap (\text{student}_{\text{SBJ}} \times W_{\text{OBJ}}))$
量化	<i>all men die</i>	$\text{man} \subset \pi_{\text{SBJ}}(\text{die})$
否定	<i>no dogs are hurt</i>	$\text{dog} \parallel \pi_{\text{OBJ}}(\text{hurt})$

図10 言語表現の表す外延の集合による抽象表現

1. $W \neq \emptyset$	5. $(A \subset B \ \& \ B \subset C) \Rightarrow A \subset C$
2. $A \cap B \subset A$	6. $(A \subset B \ \& \ A \neq \emptyset) \Rightarrow B \neq \emptyset$
3. $B \times q \subset (A, B) \subset A$	7. $(A \parallel B \ \& \ C \subset A) \Rightarrow C \parallel B$
4. $A \neq \emptyset \Leftrightarrow \pi(A) \neq \emptyset$	8. $(C \subset A \ \& \ C \subset B) \Rightarrow C \subset A \cap B$

図11 集合に関する推論の公理の抜粋

学習を行う。すなわち、

$$L(\theta) = \sum_{(x,y)} \log \sum_z p([z]=y | x, \theta) - \lambda \|\theta\|^2 \quad (9)$$

を最適化する。ここで $[z]=y$ は、 z を評価すると y になるという意味である。この方法により、論理式のアノテーションを一切必要としないにもかかわらず、質問応答の標準的なデータセットで、87.2%, 91.1%, 95.0%の驚異的な正解率が得られたことが報告されている。

ただし、DCSの意味論は充足可能性問題として定式化されているため、直接学習データに登録されていない抽象的な関係、例えば「メアリーは小説を愛読している」から「誰かに読まれている本が存在する」が含意されることを推論することはできない。これに対し最近、田らはDCSを拡張し、DCSを集合をもとにした抽象表現に写像することで、論理的推論を可能にした¹⁴⁾。この表現では、言語表現は例えば図10のように、言葉の意味する外延を集合として表す抽象表現に変換される。この上で、集合に関する図11のような公理から、言明が真であるかどうかを論理的に推論することができる。

これらは人工知能の分野で長く研究されてきた一階述語論理より制約が緩く、かつ柔軟であるために、より広い言語現象を取り扱うことができる。実際には「大合唱する」が「いっせいに異議を唱える」と同意である、といった論理的ではなく、語用論的な知識の自動獲得には機械学習が必要であり、式(7)にみられるように、この分野は論理推論と統計的機械学習のハイブリッドであるといえる。

6. 議論とむすび

映像理解にかかわると考えられる自然言語の統計的機械学習の先端について概観した。本稿では扱わなかったが、本特集で原田氏が画像について論じている方法と似た深層学習は、自然言語処理においても適用されて研究が進んで

いる。モデルが本稿で扱ったものより単純な「ニューラルネット」であり、何が学習されているのか不明であるという問題点はあるが、多くのタスクで高性能を見せており、その原理の解明と、本稿で紹介した複雑な確率モデルとの接続が研究課題となっている。

本学会誌2013年11月号の「ふぉーかす」で長島氏によっても述べられているように、映像表現は言語や音声、心理と連携する必要がある。自然言語処理においても、最近になって単純なテキストの解析から、4節で述べたような画像や動画に対する言葉のグラウンディングがようやく可能になりつつある。映像から言葉、言葉から映像という一方だけでなく、言葉と映像の両方が互いに依存する表現をモデル化することが、今後の研究に求められているのではないだろうか。

(2014年11月18日受付)

〔文 献〕

- 1) J. Lafferty, A. McCallum and F. Pereira: "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In Proc. of ICML 2001, pp.282-289 (2001)
- 2) T. Kudo: "MeCab: Yet Another Part-of-Speech and Morphological Analyzer", <http://mecab.sourceforge.net/>
- 3) D. Klein and C. Manning: "Accurate Unlexicalized Parsing", In ACL 2003, pp.423-430 (2003)
- 4) S. Goldwater, T.L. Griffiths and M. Johnson: "Interpolating Between Types and Tokens by Estimating Power-Law Generators", In NIPS 2005 (2005)
- 5) H. Shindo, Y. Miyao, A. Fujino and M. Nagata: "Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing", In ACL 2012, pp.440-448 (2012)
- 6) K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M.I. Jordan: "Matching Words and Pictures", Journal of Machine Learning Research, 3, pp.1107-1135 (2003)
- 7) 藤田早苗, 平博順, 永田昌明: "画像検索を用いた語義別画像付き辞書の構築。自然言語処理, 20, 2, pp.223-250 (2013)
- 8) D.M. Blei, A.Y. Ng and M.I. Jordan: "Latent Dirichlet Allocation", Journal of Machine Learning Research, 3, pp.993-1022 (2003)
- 9) T.L. Griffiths and M. Steyvers: "Finding scientific topics", PNAS, 101, pp.5228-5235 (2004)
- 10) H. Yu and J.M. Siskind: "Grounded Language Learning from Video Described with Sentences", In ACL 2013, pp.53-63 (2013)
- 11) Z. Ghahramani and M.I. Jordan: "Factorial Hidden Markov Models", In Advances in Neural Information Processing Systems (NIPS), 8, pp.472-478, MIT Press (1995)
- 12) L.S. Zettlemoyer and M. Collins: "Online Learning of Relaxed CCG Grammars for Parsing to Logical Form", In EMNLP-CoNLL 2007, pp.1223-1233 (2007)
- 13) P. Liang, M. Jordan and D. Klein: "Learning Dependency-Based Compositional Semantics", In ACL-HLT 2011, pp.590-599 (2011)
- 14) R. Tian, Y. Miyao and T. Matsuzaki: "Logical Inference on Dependency-based Compositional Semantics", In ACL 2014, pp.79-89 (2014)

持橋 大地 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。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、統計数理研究所准教授。統計的自然言語処理およびベイズ統計に興味を持つ。博士(理学)。