2004－SLP－50 (5)
2004／2／6

言語処理の進歩

松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

matsu@is.aist-nara.ac.jp

90 年代以降，コーパスに基づく言語処理が大きく進展した．本稿では，言語処理のタスクを問題の性質によって分類し，それぞれに対して種々の統計的機械学習が適用されていることを示す．次に，機械学習のタスクとして見た場合の言語処理の特徴について述べ，どのような性質を持つ機械学習アルゴリズムが言語処理のタスクには必要かについて論じる．具体的な言語処理の進展を見るために，英語の句構造解析と日本語の係り受け解析への機械学習の適用の変遷について紹介する．最後に，最近の機械学習の言語処理への適用の動向と展望について述べる．

キーワード: 統計的言語処理, コーパス, 機械学習, 形態素解析, 統語解析, 係り受け解析

Advances in Language Processing

Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology

matsu@is.aist-nara.ac.jp

Since 90's, remarkable advances have been observed in corpus-based natural language processing(NLP). This paper first categorizes the NLP tasks based on thier characteristics and shows various types of statistical machine learning methods having been applied to a number of different NLP tasks. Then, we describe what kinds of distinctive features in machine learning algorithims are suitable and necessary to NLP tasks. Finally, to see particular advancement in language processing, we take up parsing of English phrase structure and Japanese dependency structure and show the recent advances and trends in machine learning approaches to natural language processing.

Keywords : Statistical Language Processing, Corpus, Machine Learning, Part-of-speech Tagging, Parsing, Dependency Analyisis

1 はじめに

90 年代以降の言語処理は，コーパスの時代と呼ばれる．80 年代に単一化文法等の理論文法が盛んに研究されたのと対比して，過去十数年の言語処理の主たる進展は，大規模コーパスを用いた統計処理や機械学習に席巻されてきたと言える．著者の立場は，必ずしもコーパスに基づく言語処理一辺倒というわけではなく，詳細な文法/語彙知識との融合あるいは棲み分けを如何に考えるかという点に興味がある．しかし，近年の品詞タグ付け，固有表現抽出，係り受け解析/統語解析など種々の実用的な言語解析システムの実現は，コーパスに基づく言語処理の進展によるところが大きい．これには，次のような要因が考えられる．まず，コーパスや辞書などの言語資源を研究者間で共有化しようという動きが広まったこと，同時に WWW など電子化された言語データへのアクセスの機会が極めて急速に広まった環境面の変化が挙げられる．研究面からは，共有化されたデータの存在により，同じデータを対象に学習やシステムの

評価が行えるようになったこと，また，電子テキストデータの利用の広がりにより，広い分野を対象とする実用的な言語処理が必要とされる機会が急速に増大したことが挙げられる．

別の要因として重要なのは，近年の計算機パワーと記憶メディアの急速な進展である．90 年代前半からの過去 10 年間に，計算機のクロック数，主記憶容量とも，2 桁から 3 桁の高機能化を達成し，当時は扱うことが不可能と考えられた量のデータを日常的に扱うことができるようになった．さらに，統計的機械学習の近年の進展による寄与も大きく，大規模な素性空間を扱う必要がある種々の言語処理タスクを実用的な時間で学習，実行することができるようになってきた．

本稿では，言語処理のタスク分類と機械学習法の対応について述べ，また，どのような機能をもつ機械学習が言語処理のタスクに必要かを議論する．そして，コーパスに基づく最近の言語処理の進歩の事例として，英語と日本語の統語解析の変遷を紹介する．

2 言語処理のタスク分類

機械学習の立場から，言語処理のタスクを分類してみよう．表 1 に示すように，言語処理のタスクは，「予測」「分類」「系列タグ付け」「変換」「抽出」のように分類することができる．同表には，それぞれ対応すると考えられる具体的な処理例を示した．「予測」は，モデルとしたい対象の確率的な性質を求めるタスクであり，言語モデル，あるいは，語の確率分布の生成とそれに基づく類似度計算などが代表的なタスクである．「分類」は，最も多くの問題がその対象となるものであり，文書分類を始めとして，語義の曖昧性解消や既知のクラスへの語の分類など様々なタスクがある．どの問題も既知のクラスへの分類問題として定式化することができれば，Naive Bayes，決定木，Support Vector Machines など多種多様な分類学習法が適用可能である．

一方で，「変換」や「抽出」に対応する問題は，一般的に適用できる手法があるというよりは，個々の問題によって学習する対象が異なり，手法もそれに応じて設計されるものが多い．「変換」の代表例は，二言語の対訳コーパスにおける文，句，語などの対応付けの自動化であり，またそのような対応付けからの翻訳規則の自動抽出である．統計的機械翻訳 [Brown93] は，単語の対応や語順の対応の確率モデルを大規模な対訳コーパスから学習するもので，コーパスに基づく言語処理の先鞭となった研究である．「抽出」については，共起関係を利用した決まり文句や熟語などの抽出，クラスタリングによる意味クラスの推定，動詞の格フレームのような特定の形の語彙知識をコーパスにおける語の用法から自動抽出する研究が含まれる．また，他のタスクや学習に有効は素性あるいは素性の組み合わせを自動獲得する研究も盛んである．

共通の性質を有しつつ，単純な分類問題とは異なるタスクが「系列タグ付け (sequentila tagging)」に分類されるタスクである．この種類のタスクで最も研究されてきたのが，品詞タグ付け問題であるが，並行して統語解析の研究も大きな進展を見せている．近年では，基本句のチャンキング (base phrase chunking) や固有表現抽出 (named entity extraction) なども盛んに研究されているが，問題としては一種のタグ付け問題とみなされる．系列タグ付けの研究については，4 章で，統語解析を対象にその詳細を述べる．

コーパスに基づく言語処理の初期の研究手法は，[Manning99],[Dale00],[田中 99][Mitkov03] などによくまとめられている．

3 言語処理のための機械学習

3.1 言語処理の特徴

言語処理の基本は単語であり，一般に考慮すべき特徴量は極めて大きい．タスクによっては，単語を品詞や意味クラスに分類して特徴量を減らすことも考えられる．例えば，品詞タグ付けでは，品詞に基づく bigram や trigram に基づいたモデルを考え，単語は品詞からの生成確率の算出にだけ考えられる．初期の確率文脈自由文法による統語解析においても，句構造規則だけを対象に確率計算をするモデルが扱われていた．

しかし，品詞や句構造規則だけでは規定できない単語固有の現象が多く見られ，多くの場合，単語の個別の素性として扱うモデルも必要である．単語を素性とするような高次元空間を対象にする場合には，常にデータの過疎性 (data sparseness) 問題が起こるので，異なるレベルの素性間 (単語と品詞，意味クラスなど) のスムージングをうまく行う必要がある．さらに複雑な言語処理の問題では，幾つかの素性が同

学習タスクの分類	言語処理タスクの例
予測	言語モデル，語の類似度
分類	文書分類，語義曖昧性解消 (WSD)，用語の意味分類，照応解析
系列タグ付け	分かち書き，品詞タグ付け，基本句チャンキング，固有表現抽出，統語解析 (句構造解析，係り受け解析)
変換/対応付け	機械翻訳規則獲得，アラインメント，統計的機械翻訳
抽出/マイニング	コロケーション，クラスタリング，語彙知識獲得，素性選択

表 1: 機械学習と言語処理タスクの分類

時に成り立つことを確認するような素性の組み合わせを扱える枠組みが必要になることが多い．機械学習の問題として見た場合の言語処理の特徴をまとめると次のようになる．

離散シンボル: 品詞，意味クラス，単語，特定の句や単語列など基本とする素性が離散シンボルとして表現される．

高次元: 単語を基本素性とすることで，個々の事例が極めて高次元のデータとして表現される．

曖昧性: 一つの単語が複数の品詞や意味クラスに属する．また，語が他の語と複合することにより，異なる性質や挙動を示すようになる．

素性間の依存性: 一つの素性が，事例の所属するクラスを決めるために支配的な影響を持つ場合もあるが，統語解析のように複雑な素性の組み合わせが決定要因となるタスクがある．

不均一性: 素性の出現頻度に大きな偏りがある．データの振舞に分野および時間的依存性がある．

3.2 言語処理のための機械学習の要件

前節で述べたことから帰結されるのは，言語処理に適用される機械学習アルゴリズムには次のような性質を持つことが好ましいということである．

- 事例が離散値をもつベクトル表現で表される．文書検索における tf-idf や何らかの共起尺度による連続値をもつ場合もあるが，多くは，素性が存在するか否かの 2 値素性として表すことができる
- 極めて高次元の表現を取り扱うことができること
- 異なる粒度の素性間のスムージングを取り扱う必要があること
- 有用な素性の選択や素性同士の組み合わせを取り扱うことができること

後者の 2 点についてが，言語処理において特に重要な点である．例えば，HMM 等確率をベースにした品詞タグ付けでは，データ過疎性に対応するため，品詞と単語間，また，品詞 bigram と trigram の間のスムージングを適切に行う必要があった．最大エントロピー法に基づく [Ratnaparkhi96] の手法が品詞タグ付けによい精度を実現したのは，オーバーラップする複数の素性を明示的に指定して同時に条件部に使用し，エントロピー最大化によってそれらの間のスムージング効果が自動的に達成されたことによると言える．また，Brill の変換に基づく手法 [Brill95] が品詞タグ付けにおいてよい結果を示しているのは，比較的複雑な文脈表現テンプレートを仮定して，素性の組み合わせを直接条件付けに使用できることによる効果であると言える．

近年，Supprot Vector Machines が種々の言語処理タスクにおいてよい精度を示しているのは，極めて高次元の素性空間を扱うことができる能力よりもむしろ，サポートベクターという真に重要な事例の足し合わせによるスムージング効果と，多項式カーネル等のカーネル関数による組み合わせ素性利用の暗黙的な実現によるものと考えることができる．

4 統語解析の進展

本節では，近年の統計的機械学習に基づく言語処理の進歩の事例として英語と日本語の統語解析の進展を概観する．

4.1 英語の句構造解析

統語解析に最初に用いられた統計的技術は，確率文脈自由文法に基づく解析であった．個々の句構造

	解析モデル	再現率/精度 (%)
Magerman 1995	決定木学習	84.0/84.3
Collins 1996	単語共起確率	85.3/85.7
Ratnaparkhi 1997	最大エントロピー法	86.3/87.5
Charniak 1997	確率文法+単語共起	86.6/86.7
Collins 1999	確率+文法知識 (格フレーム,GAP, 等位節)	88.1/88.3
Charniak 2000	語彙化確率文法+最大エントロピー法	89.6/89.5
Collins 2000/2002	Voted perceptron+Reranking with Tree Kernel	89.6/89.9

表 2: 統計的手法による英語の統語解析法と精度

規則に確率値が割り当てられ，与えられた単語列を文として解析する可能な統語木のうち，木に含まれるすべての句構造規則の確率値の積が最大になる統語木が求める解析結果であるとされた．解析済みの文の集合からだけでなく，未解析の文の集合からも各句構造規則の確率値を推定するためのアルゴリズム (Inside-outside algorithm)[Charniak93] が利用された．しかし，句構造規則単位に確率値を考えるだけの方法では曖昧性解消の力に限界があり，単語の情報をも考慮しなければ精度の高い曖昧性解消が困難であることが認識されるようになった．英語の統語的曖昧性の原因として知られる前置詞の修飾先の曖昧性 (いわゆる PP-attachment 問題) を解消するために各々の句の主辞の情報に基づく統計手法が用いられ，確率文脈自由文法も，句の主辞に基づく確率値を取り入れるように発展した．

その後，[Magerman95] によって提案された決定木を用いたパーザがきっかけとなり，コーパスからの学習に基づく統語解析アルゴリズムの研究が盛んになった．Collins が単語の共起に基づく統計的パーザ [Collins96, Collins99] を提案し，単語の情報を重視した統計的統語解析法が次々に提案された．Collins の方法は，句構造木中の規則を 2 分木に分解して考え，一方を主辞として規則毎に 2 つの単語の共起確率を推定し，主辞となる単語を親へ継承していくという方法を取った．統語解析は，従来通り，確率値最大の統語木を求める形を取った．[Ratnaparkhi97] は，shift-reduce 法を 2 分化した統語解析法を想定し，基本動作の推定に最大エントロピー法を用いた方法を提案した．解析に有効と考えられる様々な素性と具体的な組み合わせを考慮し，高い解析精度を達成した．[Charniak97] は，従来の確率文脈自由文法によって絞り込まれた上位の解析木候補の中から，句構造規則の主辞や親などの広い文脈素性を条件とする確率モデルにより，解析精度向上を達成した．この研究は，これまでの history-based な解析の手法を，句構造に親やさらに上位の句ラベルを付与する parent-annotation や，句構造規則を主辞を中心に 2 分木化する Markovization などのアイデアを整理した仕事としても評価できる．[Charniak00] は，上記モデルの確率値推定に最大エントロピー法に基づく方法を採用し，さらに精度向上を達成した．Collins の最近の仕事 [Collins02] では，Tree Kernel という木全体の類似度を測る尺度を導入して，確率モデルが出力した解析木候補の順序を変更することによって，誤りを軽減する方法を示している．これらのパーザのすべては，Penn Treebank の決まった章を学習データ (chapter2–21) とテストデータ (chapter23) とし，同じデータを用いた比較を行っている．上で述べた現在までに提案された代表的な手法と精度を表 2 に示す．再現率は，正しい統語木に含まれるべき句構造がシステムによってどの程度再現されたかを示し，精度は，システムが出力した統語木中の句構造の内どの程度が正しいものであったかの比率を示している．

4.2 日本語係り受け解析

日本語では，句構造文法よりも文節係り受けに基づく統語処理が統計的解析に主として用いられている．その理由は，日本語では，句構造解析された大規模な学習データが存在しないこと，および，係り受け解析の方が特定の文法に依存することなく解析結果を共有しやすいことにあると考えられる．

単語の共起に基づいて確率値最大の係り受け解析

	解析モデル	学習コーパス (文数)	精度 (%)
藤尾 1998	単語共起+スムージング	EDR(190,000)	86.7
春野 1999	確率決定木+boosting	EDR(50,000)	85.0
内元 1999	最大エントロピー法	京大コーパス (7.956)	87.9
金山 2000	HPSG+単語統計	EDR(192,778)	88.6
工藤 2000	SVM+確率モデル	京大コーパス (7.956)	89.1
工藤 2002	SVM+階層的まとめ上げ	京大コーパス (7.956)	89.3
		京大コーパス (19.191)	90.5

表 3: 統計的手法による日本語の係り受け解析と精度

を得る方法がが藤尾ら [Fujio98] に提案された．また，最大エントロピー法による確率推定を用いた方法が内元ら [Uchimoto99] に提案された．その後、Support Vector Machine(SVM) の学習結果を疑似的な確率値として利用することによる係り受け解析が工藤ら [Kudo00] によって提案された．これらは，すべて確率モデルをベースにしており，学習法および利用する素性の違いによる精度の差が見られる．工藤ら [Kudo02] は，さらに、2 文節間の係り受けの有無を SVM に直接学習させ、段階的な文節まとめ上げを行うことで，精度を落とすことなく，より効率的な係り受け解析法を提案した．

その他，春野ら [Haruno98] が決定木とブースティングを用いた方法を提案し，金山ら [Kanayama00] は HPSG の文法規則と単語統計を用いた方法を提案している．日本語の場合は，学習データとしては，EDR コーパスの統語木か京大コーパスの係り受け解析木のデータが用いられている．これまでに提案された代表的な統計的日本語係り受け手法と精度を表 3 に示す．精度は，システムが出力した文節係り受けのどの程度が正しいものであったかの比率を表す。

5 最近の話題と展望：あとがきにかえて

HMM に代わって精度の高い確率モデルとして用いられるようになった最大エントロピー法による系列タグ付けモデル (Maximum Entropy Markov Model, MEMM) は，素性の組み合わせやオーバーラップする素性を条件付けできる特徴をもつが，個々の状態に依存する条件付き確率による次状態推定モデルのため，いわゆる label bias 問題 [Lafferty01] におちいる危険性がある．これに対し，より一般的で柔軟な条件付けを許し，かつ，事例全体の生成確率や順序を最適化する様々な方法が提案されている．例えば，[Lafferty01] による Conditional Random Field(CRF)，[Altun03]，による Hidden Markov Support Vector Machines などがその例である．また，[Collins02] の Tree Kernel は，木構造内のすべての部分木を要素とするベクトルの内積を効率的に計算するカーネルであり，木構造全体の類似度情報を学習に利用する方法である．

3 章で述べたように，有効な素性や素性の組み合わせを用いることが精度向上に必要である．また，既存の学習事例によってカバーされていない現象に対応するための有効な事例の収集を能動的に行うこと重要であり，これらの自動化が望まれる．

これに対し，一方では，個々の単語を素性としてめったやたらに導入することへ警鐘を鳴らす研究もある．[Klein03] は，parent annotaion や句構造の Markovization を慎重に行い，語のクラス分けや機能語など限定された語彙化を行うだけで，かなり高精度の統語解析が可能であることを示し，慎重な素性およびモデル選択が重要であることを主張している．

コーパスに基づく手法が言語処理のすべてを解決するはずはない．重要なのは，機械学習やマイニング等の利用による計算機パワーを活かすことによって，意味のない規則の列挙から人間が開放され，本質的に難しい問題を明らかにして行くことに機械の力を利用することである．機械学習モデルの改良や素性選択/組み合わせの自動化も，システムの精度を向上することだけが目的でなく，問題の本質を明確にするためのステップとして援用すべきである．

謝辞

本稿での議論は，奈良先端大自然言語処理学講座の機械学習とマイニングに関する勉強会のメンバーとの日頃の議論や研究成果に負うところが大である．勉強会のメンバーに深く感謝する．

参考文献

[Altun03] Altun, Y., Tsochantaridis, I. and Hofmann, T., "Hidden Markov Support Vector Machines," *Proc. ICML'03*, pp.3-10, 2003.

[Brill95] Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics*, Vol.21, No.4, pp.543-565, 1995.

[Brown93] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L., "The Mathematics of Statistical Machine Translation:Parameter Estimation," *Computational Linguistics*, Vol.19, No.2, pp.263-311, 1993.

[Charniak93] Charniak, E.: *Statistical Language Learning*, The MIT Press, 1993.

[Charniak97] Charniak, E., "Statistical Parsing with a Context-free Grammar and Word Statistics," Proc. AAAI-97/IAAI-97, pp.598-603, 1997.

[Charniak00] Charniak, E., "A Maximum-Entropy-Inspired Parser," Proc. 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp.132-139, 2000.

[Collins96] Collins, M., "A New Statistical Parser Based on Bigram Lexical Dependencies," Proc. ACL'96, pp.184-191, 1996.

[Collins99] Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. PhD Dissertation, University of Pennsylvania, 1999.

[Collins00] Collins, M., "Discriminative Reranking for Natural Language Parsing," Proc. 17th ICML, pp.175-182, 2000.

[Collins02] Collins, M. and Duffy, N., "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron," Proc. 40th ACL, pp.263-270, 2002.

[Dale00] Dale, R., Moisl, M. and Somers, H.(eds.): *Handbook of Natural Language Processing*, Marcel Dekker, 2000.

[Fujio98] Fujio, M. and Matsumoto, Y., "Japanese Dependency Structure Analysis based on Lexicalized Statistics," *Proc. 3rd EMNLP*, pp.88-96, 1998.

[Haruno98] Haruno, M., Shirai, S. and Ooyama, Y., "Using Decision Trees to Construct a Partial Parser", *Machine Learning*, Vol.34, No.1, pp.131-149, 1998.

[Kanayama00] Kanayama, H., et al., "A Hybrid Japanese Parser with Hand-crafted Grammer and Statistics," *Proc. COLING2000*, Vol.1, pp.411-417, 2000.

[Klein03] Klein, D. and Manning, C.D., "Accurate Unlexicalized Parsing," *Proc. ACL'03*, pp.423-430, 2003.

[Kudo00] Kudo, T. and Matsumoto, Y., "Japanese Dependency Structure Analysis Based on Support Vector Machines," *Proc. EMNLP-VLC-2000*, pp.18-25, 2000.

[Kudo02] Kudo, T. and Matsumoto, Y., "Japanese Dependency Analysis using Cascaded Chunking," *Proc. CoNLL-2002*, pp.63-69, 2002.

[Lafferty01] Lafferty, J., McCallum, A. and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proc. ICML'01*, pp.282-289, 2001.

[Magerman95] Magerman, D.M., "Statistical Decision-Tree Models for Parsing," *Proc. ACL-95*, pp.276-283, 1995.

[Manning99] Manning, C.D. and Sch"utze, H.: *Foundations of Statistical Natural Language Processing* The MIT Press, 1999.

[Mitkov03] Mitkov, R.(ed.): *The Oxford Handbook of Computational Linguistics*, Oxford Univ Press, 2003.

[Ratnaparkhi96] Ratnaparkhi, A., "A Maximum Entropy Model for Part-of-Speech Tagging," *Proc. the Conference on Empirical Methods in Natural Language Processing*, pp.133-142, 1996.

[Ratnaparkhi97] Ratnaparkhi, A., "A Linear Observed Time Statistical Parser Based on Maximum Entropy Models," Proc. 2nd Conference on Empirical Methods in Natural Language Processing, pp.1-10, 1997.

[田中 99] 田中穂積 (ed.): 自然言語処理–基礎と応用, 電子情報通信学会, 1999.

[Uchimoto99] Uchimoto, K., et al., "Japanese Dependency Structure Analysis Based on Maximum Entropy Models," Proc. 9th Conference of the European Chapter of ACL, pp.196-203, 1999.