6 コーパスに基づく言語処理の限界と展望

松本 裕治 matsu@is.aist-nara.ac.jp
奈良先端科学技術大学院大学情報科学研究科

徳永 健伸 take@cl.cs.titech.ac.jp
東京工業大学大学院情報理工学研究科

1990年代に入って盛んになった「コーパスに基づく言語処理」の研究は，過去10年で処理のほとんどあらゆる分野に浸透した．現在では，コーパスに基づく手法は，重要な方法論として言語処理の一大分野をなしているといえる．

本稿では，まずその背景，適用範囲，現状について概観する．その後，コーパスに基づく言語処理の得失と評価における問題点について述べ，今後の展望について論じる．

コーパスに基づく言語処理概観

1990年代に入ると同時にコーパスの利用が大きく進展したのには次のような要因がある．1つは，電子化された大規模な言語データが計算機上で利用可能になってきたことである．米国では，ペンシルバニア大学のLDC（Linguistic Data Consortium）がWall Street Journal等のテキストを自動的に解析し，人手で修正することにより品詞情報や構文情報などのタグを付与した，いわゆる解析済みコーパス（タグ付きコーパス）を提供しはじめた．これにより，共通のデータを利用して客観的な研究を行うことが可能になった．日本でも，1980年代半ばから開始された電子化辞書プロジェクト（EDR）の成果が1990年代に入って公開され，一般に利用できるタグ付きコーパスとなった．また，日経新聞や毎日新聞などの全記事を収録したCD-ROMが研究用として安価に利用できるようになった．

2つ目の要因は，計算機の性能（速度，記憶容量）の急速な向上であり，データが利用可能になるのと符合してメガバイトあるいはギガバイト規模の言語データを扱えるようになったことである．ワークステーションの主記憶が数メガバイトであることが当り前であった1980年代には取り扱うことが困難であった規模のテキストデータや電子化辞書が今では廉価なパーソナルコンピュータでも簡単に扱うことができる．

3つ目の要因として，1980年代までの規則あるいは制約に基づく言語処理の限界が感じられ，ある種の閉塞感が漂っていた状況において，隠れマルコフモデル（HMM）などの技術を利用した品詞タグ付けシステムの提案や対訳テキストを用いた統計モデルによる機械翻訳などの仕事が華々しく登場してきたことにある．

その後，コーパスに基づく言語処理は，表**-1**のような多くの分野に適用されるに至っている．具体的な研究内容については，初期の特集論文[1]や最近の刊行物[3), 4)]を参照されたい．

コーパスといってもさまざまな形態のものがあり，用途も多岐にわたる．どのようなコーパスを，どのような目的で，どのようにして用いるかを明確にする必要がある．コーパスには，未加工のテキストデータか，ある程度解析を行ったタグ付きコーパスかの区別

言語解析システム	形態素解析（分かち書き，品詞タグ付け）， 名詞句のまとめ上げ，確率パーザ）
曖昧性解消	語の多義性，係り受けの曖昧性解消
語彙知識獲得	動詞の格フレームと選択制限， 同義語，シソーラス構築，語のクラスタリング
パターン抽出	共起，コロケーション，熟語表現，固有表現， 専門用語等の自動抽出
機械翻訳	統計的機械翻訳，事例に基づく機械翻訳， 対訳文のアラインメント，対訳表現抽出
文書処理	情報検索，情報抽出，文書クラスタリング， 文書分類，質問応答，文書構造解析， 照応解析

表-1　コーパスに基づく言語処理の分類

がある．使用目的は表-1で示したようにさまざまなものがある．また，学習のためのアルゴリズムは，N-gramやHMMなどの確率モデルのように用途に特化したものから決定木，決定リスト，最大エントロピ法のように一般的な手法などあらゆる学習アルゴリズムが対象となっている．

コーパスに基づく言語処理の得失

コーパスに基づく方法の優れた点として，計算機のパワーを活かした学習の自動化，および，大量のデータを利用することによる適用範囲の広さと頑健さが挙げられる．学習法や学習に用いる具体的な情報の種類や粒度（学習モデル）は人間が考える必要があるが，データの追加や学習モデルの変更を行いつつ学習を繰り返すことが可能であり，着実な進展を期待できる．実際，表-1に挙げた多くの分野で，既存の規則に基づく手法より高い精度の言語処理が可能になっている．また，具体的かつ共通のデータを用いることで，客観的な評価を行うことができる．これは技術の健全な進歩のために重要である．なお，評価については，注意すべき問題があり，次章で取り上げる．

一方，欠点として注意しておかねばならないのは，言語処理のタスクを限定し，タグ付きデータによって評価することにより，評価の対象以外への注意を向けなくなることである．システムの性能を数値で測ることにより，数値の向上に大きく寄与しない言語現象が無視されがちになる．言語学が関心を持つさまざまな言語現象は，ある意味で特殊な現象であり，新聞記事などのよく利用されるデータにはほとんど出現しないものが多い．そのため，研究の対象が偏ったものになる．同様に，どのような大規模データを集めても，きわめてまれにしか見つからない，あるいは，まったく見られない言語現象が必ずあり，そのような現象には当然対処することができない（データの過疎性の問題）．

評価尺度のいろいろ

現在行われている自然言語処理システムの評価は，正解情報を付与したコーパスをあらかじめ用意しておき，評価するシステムの出力と正解を比較することによって行うのが一般的である．正解情報は人手で用意しておく必要がある．評価尺度は解析対象となる解析レベル（形態素，構文，意味など）によって異なるが，一般的に用いられる尺度として再現率（recall）と精度（precision）がある．再現率は正解に含まれる情報をシステムの出力がどれだけ網羅しているか，精度はシステムの出力がどれだけ正しいかを表す尺度である．

たとえば，日本語のように語の境界を明示しない言語の形態素解析では，まず語の境界を同定する必要がある（分かち書き）．語境界の同定の再現率はシステムが出力した正しい語境界の数を正解コーパス中の語境界の数で割った値で定義される．また，精度は，システムが出力した正しい語境界の数をシステムが出力したすべての語境界の数で割った値となる．形態素解析は，さらに各語に正しく品詞情報を付与する必要がある．

分かち書きを対象にするか単語の同定と品詞付与を対象にするかによって，これらの尺度が異なる様子を，次の例文で見てみよう．(1) が正解，(2) がシステムの出力とする．また，スラッシュを単語の区切り

とする．

(1) ／学校／に／はいっ／た／．／
(2) ／学校／に／は／いっ／た／．／

分かち書きとしてこれらの文を見ると，(1) では (文頭と文末を除いて) 4カ所の語境界があり，一方 (2) は5カ所の語境界がある．これらのうち4カ所が一致している．よって，再現率は 4/4 = 100%，精度は 4/5 = 80%である．ところが，単語の出力の評価を行うと，(1) では5単語，(2) では6単語が得られており，そのうち，4単語が一致している．よって，再現率は 4/5 = 80%，精度は 4/6 = 67%となる．品詞情報まで考慮する場合は，各語の品詞名まで一致して初めて正解とみなされる．

句構造文法を基礎とした構文解析システムの評価で現在よく使われているのは正解コーパスの括弧付けとシステムの出力の括弧付けを比較する方法である．括弧の数に関する再現率と精度，あるいは正解コーパスの括弧とシステムの出力の括弧が交差する数 (crossing) によってシステムの性能を評価する．この手法では，構成素のラベルを無視するので，正解コーパスで用いられている文法と解析システムの文法の間で構成素ラベルが異なっていても評価ができる．いいかえれば正解コーパスには括弧付けだけがあればよいという利点がある．しかし，交差数に関しては括弧の構造的な位置によって交差数が大きく影響を受けるという欠点が指摘されている．

これに対して依存文法を基礎とした構文解析システムの評価では，いわゆる係り受けの数に関する再現率と精度を利用することが多い．Carroll らは依存関係に主語 - 述語関係，修飾関係などの文法的関係を付加した2項関係を性能評価の基礎に使う評価方法を提案している[2]．

正解コーパスに基づく評価について，いくつかの問題点を指摘しておきたい．上記の形態素解析や構文解析の評価尺度からも分かる通り，何を正解と考えるかによって，同様の処理に対しても精度等の数値は異なる．特に，品詞付与については，同一言語に対しても品詞体系が数多く存在するという問題がある．

たとえば，英語の解析における標準的なコーパスとなっている Penn Treebank では，45種類の品詞からなる体系が用いられているのに対し，その体系の基礎となった Brown コーパスでは87種類の品詞が用いられていた．また，英国の ICE (International Corpus of English) コーパスでは，約20種類の品詞大分類の下に，人称，数，時制などの素性を持たせており，すべてを展開すると数100種類の品詞を持つことになる．日本語においても，著者の1人がかかわっている茶筌[5]という形態素解析システムでは，RWCP コーパスで用いられている IPA 品詞体系を利用しており，これは4階層の品詞体系からなる．用言の活用型や活用形をすべて展開すると，約500種類の品詞を持つことになる．品詞タグ付けの正解をどの細かさで判定するかによって再現率や精度の数値が大きく変わることは明らかであろう．タグ情報の詳細さを考慮せずに数値のみを比較するのは危険である．

他に無視できない重要な問題として，正解コーパスの中に残っているエラーの存在がある．たとえ人手でチェックされたものであっても，数万文あるいは数百万語からなるデータには，必ず，見逃しによるエラーや人為的なミスによるエラーが含まれてしまう．

多くの評価方法は，解析システムの一般的な性能を評価するために提案されているものである．したがって，解析システムを具体的な応用システムに利用する場合には，その応用システムの特性を考慮しないと，一般的な評価による性能がよいからといって必ずしも期待する結果が得られない可能性がある．筆者の1人は構文解析を情報検索の索引付けに利用する研究を行っている．予備的な実験として，解析システムとして NYU で開発された Apple Pie Parser と CMU で開発された Link parser を用い，構文解析されたドキュメントから上述した Carroll らの提案している文法関係を抽出し，それを索引語として利用し検索を行った．文法関係を正確に抽出する性能は，Link parser の方が Apple Pie Parser よりもよかったが，情報検索の性能としては結果が逆転した．この

原因を分析した結果，索引語として有効な名詞句の解析性能がApple Pie Parserの方が優れていることが分かった．

また，ドキュメントのタイトルには有効な索引語が含まれていることが多いが，タイトルの多くはいわゆる「文」の構造をなしていないために，Link parserでは解析に失敗することが多い点も原因の1つである．このように，一般的な性能評価だけではなく，応用システムに自然言語処理の技術を用いる場合は，その応用独自の観点から解析システムを評価する必要がある．

コーパスで重要なのは質か量かという議論がある．質の高い，すなわち，詳細なタグを持ったコーパスの作成には多大なコストがかかり，大規模なコーパスを作成することが難しい．一方，タグ付けを考えなければ，電子的なテキストデータは，今や（種類は限られるものの）膨大な量が入手可能である．

HMMなどの学習法や共起などの表層的な情報は，未加工のデータからある程度の情報を得ることができるが，獲得すべき言語情報が深くなるほど，質の高いタグ付きコーパスが必要になる．ただし，統計的な解析技術の精度が向上することで，その結果を利用した格フレーム獲得の精度が上がるように，ブートストラップ的な考え方によって未加工のデータを利用することが考えられる．正解データに基づいて行う教師付き（supervised）学習に対して，未加工のデータに基づく学習を教師無し（unsupervised）というが，教師無し学習であっても，精度の高い一部の学習データをブートストラップ的に使うことによって未加工のデータを有効に使う方法も検討されている．

いかに人間の手間を一部の
難しいデータの作成のみに
集約させるか？？？

今後の進展への期待

コーパスに基づく言語処理は，技術として定着したといえる．その一方で，より深い言語処理を行うには，より詳細なタグ付きコーパスの構築が必要であり，これは容易なことではない．さらに，タグの種類が複雑になればなるほど，エラーが入り込む余地が増えることになる．タグ付きコーパスの蓄積を整合的に行うために，いかにコーパスおよび学習手法を利用するかといったコーパス構築支援が重要な課題である．この場合にも，前章の最後に述べたブートストラップ的な考え方が大切であろう．また，人手で構築された辞書やシソーラスなど既存の言語資源を土台にして，それを拡張したり，新しい領域の文書の解析に適用したりする意味でのブートストラップも重要である．これは広く捉えると，人間の知見によって記述された知識とコーパスから取り出された知識をいかに融合するかという方向であり，両者の技術の相補的な利用を模索することが重要であろう．

有効な言語知識の学習のためにタグ付きコーパスが必要であるが，量的な改善は学習結果の質的改善に比例するのではなく，その効果は急激に減衰する．すなわち，量的に0から100まで構築したタグ付きコーパスの効果は大きいが，100から200に量を増やしても効果ははるかに少なく，同様の効果を得るには，その10倍あるいは100倍のデータを必要とする場合がある．したがって，大規模なコーパスの作成を単調に行うことはすべきではなく，いかに未加工のデータを有効に利用するか（たとえば，教師無し学習），および，いかに有効なデータを同定し，人間の手間を一部の難しいデータの作成のみに集約させるかという視点を持つことが重要である．

参考文献

1) Armstrong, S. (ed.): Using Large Corpora, The MIT Press (1994). Also in Computational Linguistics, Special Issue on Using Large Corpora, I, II, Vol.19, No.1, 2 (1993).
2) Carroll, J., Briscoe, E. and Sanfilippo, A.: Parser Evaluation: A Survey and a New Proposal, Proc. the 1st International Conference on Language Resource and Evaluation, pp.447-454 (1998).
3) 北　研二: 確率的言語モデル, 言語と計算, Vol.4, 東京大学出版会 (1999).
4) Manning, C.D. and Schütze, H.: Foundations of Statistical Natural Language Processing, The MIT Press (1999).
5) 松本裕治, 他: 日本語形態素解析システム『茶筌』version 2.0使用説明書, 第2版, 奈良先端科学技術大学院大学, NAIST-IS-TR99012 (1999). http://cl.aist-nara.ac.jp/lab/nlt/chasen/

（平成12年5月24日受付）