巻頭言，自然言語処理，Vol.17，No.1，2010

問題を語る

コーパスベースと呼ばれた統計的言語処理のうねりが誰の目にも明らかになった 1990 年代初頭からもう 20 年になる．昨年出版された言語処理学事典を広げると，この間の歴史が一望できて興味深い．たとえば，この 20 年で大きく発展した領域がある一方，当時からほとんど変わっていない領域もあることに改めて気づかされる．

進んだ領域とそうでない領域，両者を分けた要因はいくつもあるだろう．統計的機械学習の目覚ましい発展を直接享受できた領域は強かったし，ウェブのような大規模言語データの出現は知識の自動獲得に現実味を与えた．一方，談話解析や意図理解のような領域がなかなか進まないのはそもそも問題が難しいから，という議論もあるだろう．いずれにも異論はないが，見過ごすべきでない要因がもう一つある．ベンチマークデータ，そして問題設計の有無である．

ベンチマークがあれば，手法を横並びに評価でき，本質的に重要な部分が明らかになって，技術の一般化が進む．Penn Treebank や京都コーパスが形態素・構文解析の評価用標準データとして繰り返し使われてきたのはその好例だ．また，ベンチマークは問題を規定する．何を固有表現と認めるのか，述語の項を何種類に分類するか，照応関係の認定条件をどうするかなど，問題の具体的な取り決めは主としてベンチマークの設計者が行ってきた．ベンチマークによって問題が規定され，研究者間で共有されれば，技術は進む．コーパスベース，すなわち経験主義的アプローチが拡大したこの 20 年は，ベンチマークの開発による問題設計の歴史としてもたどることができる．

これらの問題設計の多くが具体的な応用の課題分析から始まっている点にも注目したい．振り返ると，固有表現の重要性が広く認識されるようになったのは，米国の評価型研究プログラム MUC の情報抽出タスクがきっかけだった．情報抽出の研究プログラム ACE からは，entity と mention を区別した共参照解析の課題仕様が提案され，意見分析や評判抽出と呼ばれるウェブ時代のアプリケーションは，subjectivity や sentiment polarity といった新しい観点の意味解析課題を生んだ．情報抽出や質問応答の中心的な問題が言語表現間の同義・含意関係を認識する問題に帰着することから，これを応用横断的な基本問題として切り出し，独立に開発・評価する動きも進んでいる．こうして問題は発見され，創造され，修正されてきた．具体的な応用から出発して，実際の言語データと格闘しながら問題を練っていく．問題の切り方が良ければ，やがては応用横断的な基本問題に一般化され，技術のモジュール性，そして統合のしやすさにも繋がる．初歩的な確認にすぎないが，問題を設計することの貢献は大きい．

ひるがえって，スクリプトの認識やプランの推定，意図理解，対話理解といった，コーパスベース以前の時代には盛んに論じられ，深い言語理解に繋がると思われた話はどうだろう．これらの領域の技術が進まないのはまだ誰もうまい問題の設計に成功していないからだ，というのは言い過ぎだろうか？ もちろん一つには知識のボトルネックの問題があっただろう．しかし，大規模コーパスからの多様な知識獲得が現実味を帯びてきた今，知識不足の障害は一気にではないにせよ，これまで経験したことのない程度に解消される可能性が出てきた．とすれば今こそ，ウェブで広がる応用を見据えながら問題の仕立て直しを検討し，その設計に知恵を絞るべき時期ではないか．

言語の理解という目標はまだ遠く漠然としており，眼前には茫洋たる空間が広がっている．そこにどんな道筋を描くか，どうすれば意味の問題をうまく分解できるか，どんな応用が広がるか，語るべき問題はいくらでもあり，いよいよ佳境．そういう時機に立ち会えることを幸運に思う．

乾健太郎（奈良先端科学技術大学院大学）