曖昧になる境界、散り散りになる課題

乾 健太郎
東北大学大学院情報科学研究科
inui@ecei.tohoku.ac.jp

研究領域の境界は本来的に曖昧である。ここまでが領域 A で、ここから先は B、のように明確に線を引くことはできない。この曖昧さは領域の成長や周囲の環境に応じて変化もする。境界が鮮明になり硬化する時期もあれば、不鮮明になり軟化する時期もある。その意味で自然言語処理は今、境界が大きく不鮮明化する時代を迎えているように見える。

たとえば、自然言語の解析と画像・動画の解析の相互乗り入れが近年急速に進んでおり、今後も加速すると予想される。社会のサイバーフィジカル化が進み、実世界の多様なデータが入手可能になると、言語という記号の世界に閉じていた処理を実世界の個体や時空間に紐付けるグラウンディングがますます重要になるだろう。コミュニケーションの媒体が多様化する中、研究の対象はバーバルコミュニケーションからノンバーバルコミュニケーションに大きく広がっていくと思われる。テキストマイニングを含むデータ解析では、言語処理（機械処理）とクラウドソーシングの境界が曖昧になりつつあるし、言語処理単体の性能よりもむしろ他の技術や領域との組み合わせ方の設計が問われるようになる。

このように、自然言語処理は多方面に開かれた、鮮明な境界を持たない研究領域になりつつあるのかもしれない。

これに関連して、確認しておいた方がよいと思われる兆候がもう一つある。言語の意味の問題に立ち入る深い処理へと研究が進むにしたがって、捉えるべき言語現象の裾野が加速的に広がっているという問題である。我々の研究分野は多様で散り散りの個別的な課題を相手にしなければならないフェーズにさしかかっているのかもしれない。

意味の問題については、関根聡氏（楽天技術研究所・ニューヨーク大学）が本ワークショップ予稿の中で極めて示唆深い指摘を行っている。

> 照応解析、情報抽出、情報検索、対話処理などの問題になると、文字というシンボルの操作だけでは 60%程度の精度を超えることが極端に難しくなる。（中略）なぜ、10 年くらいの研究で 60%に到達した後に下火になるのであろうか？個人的には、ここには「意味」という問題を内包した殻の固い卵があり、それを壊せずに衰退しているという状況だと考えている。そして、それらの卵はすべてが「意味」という同じ卵なのではないかと考えている。

精度「60%」の向こうに「意味」の未解決問題が待っていることに異論はないだろう。大学はそこに進むべきという関根氏からの熱いエールにも大学人の一人として応えられるようになりたいと願う。

意味の問題に踏み込む研究は、関根氏も言うように、簡単ではない。ただし、その最大の原因は、「「意味」という同じ卵」の中にあるというよりはむしろ、「60%」の先の課題があまりに多様で個別的なことにあるのではないかという印象を持っている。照応解析や含意関係認識などで実際の解析誤りを見てみると、何か一つの根源的な問題がそこにあって、それが解ければ一気に精度が上がるという構造になっている訳ではどうもなさそうである。「60%」の先には、オントロジカルな語彙知識や因果関係・スクリプトなどの世界知識が関わるすぐに思い浮かぶ類いの問題の他に、実に多様な問題が少しずつ顔を見せる。数量の計算の問題、時空間の推論の問題、領域ごとの慣習的スタイ

ル、知覚にからむ問題、オノマトペ、メタファーなど、雑多な問題が少しずつ混じっていて、それらの誤りの蓄積で精度が伸びない。それぞれはおそらくかなり違う解法が必要である。しかし、どれか一つにアドレスして解法を作っても、それに関わる現象は元の照応解析や含意関係認識の評価用データにわずかしか含まれないし、正解データにはノイズも混じっているので、全体の精度への寄与を定量的に測定するのは大抵の場合極めて困難である。したがって、なかなか元気が出ない。また、特定の問題だけを含むようなデータセットを作ろうとしても、不自然なサンプルになることが多く、今のところあまりうまくいっていない。

このように、「60%」の先の問題はかなり性質の違う雑多な課題の集まり・組み合わせになっており、それぞれを個別に解決しようとしても、その効果を測定する自然なデータセットを作るのが難しいのでなかなか研究が進まないという状況があるのではないか。最初に述べた自然言語処理の境界の融解がその傾向に拍車をかける可能性もある。言語の意味の問題に踏み込んでいくことは重要である。だからこそ、なおのこと、何がなぜ難しいのかをさらに深く理解していく努力とそれを踏まえた研究の方法論の議論が求められているように思われる。