言語処理、言語学、そして「つぶやき」

田中穂積
東京工業大学大学院情報理工学研究科

1　はじめに

言語処理に関する研究がさまざまな学会（情報処理学会、電子通信学会など）で分散して発表されていたこと、それらを分散させず、まとめて論文発表する場を作ることが、我が国における言語処理研究の健全な発展にとって好ましいことではないか、規模は小さくとも文系理系を問わず、言語処理を巡って議論する場を設けることが必要ではないか、という議論が言語処理学会設立の背景にあったかと思う。言語処理学会は、設立当初のもくろみどおり文理融合の学会として成長してきたといえるだろうか。

数年前の言語処理学会年次大会で、言語処理を巡る文理融合の立場から、パネルが開催されたことがある。そのころ、最近の言語処理学会の研究発表は、コーパスベースあるいは統計ベースの発表に偏り過ぎていて、これで「良いのだろうか」というつぶやきを何人かの研究者から聞いていたので、タイムリーな企画であると思った。おそらく、「これで良いのか」というつぶやきには、文系理系を問わず、コーパスベースあるいは統計ベースの研究だけでは、言語の本質には迫れないのではないかという懸念が込められていたように思う。この問題を少し考えてみたい。

2　すれ違い

我々は、母国語であれば何の苦もなく文を解釈し意味を理解することができる。通常、自然言語は、常識や前後の脈絡や意味を考えて、ほとんど唯一つの解釈に絞り込む能力を我々がもっているからである。ところがコンピュータで（自然）言語を処理をしたことがある人は、一つの文から多数の処理結果が得られることを知り、その後の処理に途方にくれたといった経験をしたに違いない。長文になれば、処理結果の数が何百万、何千万、ときには天文学的な数字になることもまれではない。その中から意味的にも文法的にも妥当な処理結果（普通は数個）を取り出すことは、1 km 先の針の穴に糸を通すような至難の技に近いように見える。これは言語処理における曖昧性解消の問題とよばれている。曖昧性解消の問題は、言語処理の研究者の前に立ちはだかる最大の問題であった。そこで、多数の言語処理結果に何らかのスコア付けすることが新たな課題として浮上する。言語処理結果にスコア付けすることができれば、スコア順に数個の処理結果を取り出すことにより、曖昧性を解消することができるからである。

次の問題はコンピュータに（言語処理モジュールに）、このスコアをどう計算させるかということである。現在の言語処理技術のレベルは、処理すべき文の前後の脈絡や意味を考えることが苦手である。ましてや人間のように常識を使った文の自在な解釈などできるレベルにはない。このレベルでどう言語処理結果にスコアづけをしたら良いのか。現在の言語処理の研究者がたどり着いた一つの結論は、大量の文例を集めて、そこからたとえば修飾関係についての統計的情報を引きだし、それを曖昧性解消のための確率的なスコアの計算に利用しようということであった。確率値は 0 から 1 の間で正規化されているため、比較の基準となるスコアとして極めて望ましい性質をもっている。しかも、言語処理技術が最も苦手とする、文の前後の脈絡や意味を考えなくて済むという利点もあった。言語処理の研究の多くがコーパスベース、統計ベースの枠組に惹かれていった背景には、このうした事情も関係していたと推察される。

言語処理より一足早く音声認識の分野では、コーパスベース、統計ベースの研究が技術のブレイクスルーをもたらしていた。雑音下での音声認識など幾つかの問題が残されてはいるが、コーパスベース、統計ベースの技術により、音声認識が実用の域に達していた。適用は遅れたが、言語処理

の分野でも、コーパスベース、統計ベースの技術は確かに目覚しい成果を上げ、これまでの言語処理の水準を一気に向上させたといってよい。そして、電子化された多量の文例集（コーパス）を集めることが、言語処理技術の基盤として重要であるという認識も高まってきた。

その間、言語処理研究者の悪戦苦闘の歴史ともいえる曖昧性解消の問題を、言語学者もまったく論じていなかったという訳ではない。典型的には英語の前置詞句付加の問題がある。”I saw a boy with a telegraph” を取り上げる。前置詞句” with a telegraph” が直前の名詞句”a boy” を修飾する解釈より、動詞句” saw a boy” を修飾する解釈の方が優先するかどうかとか、この両者に優劣がつかず、脈絡を考えなければ解消できない曖昧性がある、というような議論はなされている。いずれも曖昧性解消の問題を、人間の言語理解能力の問題に帰着させた議論であり、コンピュータでこれを実現することは難しい。前置詞句が、直前の名詞句を修飾しやすいか動詞句を修飾しやすいかという傾向についての議論もなされてはいるが、大量の文例を集めて、それを調べて一歩先に踏み込むレベルには至らなかった。

曖昧性解消の問題を解決しない限り、言語処理技術の向上は見込めないとする言語処理の研究者に対して、言語学の研究者は、それは言語学の”one of them” の問題であると見なしていた。このあたりの感覚的な「ずれ」が、言語処理と言語学の研究者との間に「すれ違い」を生んだように思う。おそらくこの「すれ違い」が、研究発表がコーパスベースや統計ベースの研究に偏重していた頃に、これで「良いのか」という言語学者からの「つぶやき」を生んだのではないか。

3 「つぶやき」は「つぶやき」として「その先は？」

前節で述べた言語学者らの「つぶやき」を単刀直入に言えば、「単語の生起頻度を測るだけで言語の本質に迫れるか」という問いで置き換えることもできるかもしれない。実は当時言語処理の側でも、このような「つぶやき」をもらす研究者がいたと思う。最近では、この「つぶやき」は、華々しいコーパスベースと統計ベースの研究成果の陰に埋もれてしまったように見える。確かに現在の言語処理技術のレベルアップに、コーパスベース、統計ベースのアプローチは確実に貢献したと断言できる。それを支えたインフラとして、大規模なコーパスの構築があった。

コーパスベースと統計ベースの研究手法は、マイナーな改良はこれからあるにしても、主な手法はほぼ出尽くしたと見て良いだろうか。研究成果にしても、ほぼ飽和してきたと見て良いだろうか。研究成果の飽和現象は、コーパスの量の不足に帰着させる考え方もあろう。しかし、この飽和現象は、そうではなさそうだと言う気がする。よりよい成果を上げるために、どのパラメータをどう操作したら良いかが判然としない不透明なモデルであることにも起因しているような気がする。言語処理の統計的に処理する部分を暗箱とみなして、そこから答を得ても、その答の意味を問うことが難しいからである。統計的な手法は、なんらかの数値が答として得られるので、深い言語現象の分析がなおざりになりがちであることも問題かもしれない。筆者は、現在は、主要な手法がほぼ出尽くして、得られる成果も飽和してきていると見ているがどうだろうか。

そこでもう一度、言語学者の「つぶやき」の意味を問い直すことも意義あることではないだろうか。曖昧性の解消には、我々は明らかにシンタックスだけでなく意味や脈絡を使っている。曖昧性解消に常識を用いていることもある。意味や脈絡や常識まで総動員して曖昧性を解消していると思われる。残念ながら言語処理、言語学の立場から、この困難な問題に立ち向かう研究者の数はさほど多くない。人間がどのような機構で言語を理解しているかということは、認知科学が扱うべき最大の問題であろう。これはまた言語処理の研究者がこれから挑戦すべき課題の一つであり、この問題を解決することが「その先は？」に答えることにつながるだろう。そのために今何をなすべきか。その答えは、言語処理と言語学の研究者が協力し、もっと議論し合う場と機会を設けることである。その意味で、これから言語処理学会の果たす役割はますます重要になると思う。ここまで書いてきてページが尽きたことに気が付いた。「その先は？」の具体的な答えはこのパネルで、あるいはパネルの後で会員同士でさらに議論していただけたらと思う。