

# 言語処理の過去，現在，そして今後

徳永 健伸

東京工業大学 大学院情報理工学研究科

take@cl.cs.titech.ac.jp

## 過去

機械翻訳 (MT) は言語処理の代表的な応用であるが，これは計算機の発明とほぼ同時に考えられた応用のひとつであった (Hutchins, 2000)．特に 1957 年のスプートニク・ショックのために露英翻訳の需要が高まり，計算機を使って強引に自動翻訳を実現することが試みられたが，この試みはほどなく挫折する．ほどなく出た ALPAC レポート (Pierce et al., 1966) により，計算言語学の基礎研究の重要性が認識されるようになる．この方向転換は ACL の前身の AMTCL (Association for Machine Translation and Computational Linguistics) が 1968 年に ACL に改名されたことに象徴されている．ここから 1980 年台まで，しばらく MT の不遇の時代が続くが，この間，言語処理の応用として SHRDLU (Winograd, 1972) や LUNAR (Woods, 1978) に代表される対話 (QA) システムが研究されていたことは興味深い．1980 年代になって MT の研究が再び活発になり，大規模な知識の必要性が認識されるようになった．当時は人手で知識を構築していたが，1990 年代に入り電子化テキストの普及，さらには Web の爆発的な発展によって，大規模な言語データが利用可能になり，現在に至るデータ駆動の研究アプローチが始まる．

## 現在

現在のデータ駆動の研究を支えているのは，インターネット上から入手可能な大規模な言語データと機械学習技術のめざましい進展であろう．現在は良くも悪くも言語処理の研究を進める上でデータは欠かせない．データを入手したら問題を分析あるいは設定し，データに対して種々の情報をアノテーションすることによって訓練・評価データを作成する．さらにそのデータに種々の機械学習の手法を適用し，処理結果を得る．これを種々の評価尺度によって評価し，さらに学習に使う素性を改良したり，パラメタを調整することによって改良する．この過程は Pustejovsky and Stubbs (2012) によって MATTER サイクル (Model, Annotate, Train, Test, Evaluate, Revise) として定式化されている．

## 今後

現在のデータ駆動の研究アプローチは，形態素解析や統語解析など，問題が比較的明確に定式化されている基盤技術においてめざましい性能の向上をもたらした．これらの基盤技術をツール化することに成功している．これは意味や談話処理のような次の段階の研究課題に容易に取り組む環境を整えたという点で言語処理の分野に対する大きな貢献である．しかし，一方で，問題の定式化が不十分な分野にも，その有効性を十分に吟味せずに同じアプローチを適用したり，何かを解明したいからデータを分析するのではなく，データがあるから適当な問題を設定して，後は性能を上げるためにひたすら「どうやるか」に血道を上げるという弊害も生んでいるように思える．言語本来の機能がコミュニケーションであることを再認識し，言語学や認知科学などの関連分野との係りをもっと深めた方向の研究が増えることが必要ではなかろうか (Reiter, 2007; Church, 2011)．

## 参考文献

- Kenneth Church. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6(5):1–27.
- John Hutchins. 2000. *Early Years in Machine Translation*. John Benjamins.
- John R. Pierce, John Carroll, Eric Hamp, David Hays, Charles Hockett, Anthony Oettinger, and Alan Perlis. 1966. Language and Machines: Computers in Translation and Linguistics. Technical report, National Academy of Sciences, National Research Council.
- James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly.
- Ehud Reiter. 2007. The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287.
- Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.
- William A. Woods. 1978. Semantics and quantification in natural language question answering. In Marshall C. Yovits, editor, *Advances in Computers*, volume 17, pages 2–87. Academic Press.