自然言語処理の歴史的変遷

参考：辻井潤一「ことばとコンピュータ」月間言語に2000年に連載

言語論の歴史を振り返ると:
古代編

I. ロゴス:あらゆる話し言葉の根底にあって、それに生命を与えている理性的能力
II. 古代ギリシアにおける言語研究(完成度の高かったギリシア語)
    I. 言語は変化する。年を経るうちに見失われた真の意味を求める学
    II. 議論された問題は
        I. 言語は自然の基づくのか、慣習に基づくのか
        II. 言語は規則性を根本原理として成り立っているのか
        III. 品詞はいくつあるのか
III. モノには正しい名前がある:ソクラテス
IV. 言語の背後の論理へ:アリストテレス
V. 修辞法の習得へ:クインティリアヌス
    I. 技能の階層:文法学、論理学、修辞学

- 話言葉から書き言葉へ
- 観念から実用への流れ

言語論の歴史を振り返ると
中世編

I. 1000年以上にわたってラテン語がヨーロッパの共通言語であり続けた。
   I. Realist=普遍語（人、馬など）は実体を持ち、物理的実体に先立つ
   II. Nominalst=個々の事物が実体であり、普遍語は単なる抽象物（記号）である
II. 1453年のコンスタンチノポリス陥落
   I. ラテン語学者たちがイタリアに戻る
   II. ギリシア、ローマの古典の復興
   III. しかし、ヨーロッパは分裂し、中央集権国家は、土着の言語を国家言語として利用し、国家をまとめた。→ラテン語の衰退
   IV. 経済のグローバル化、技術の発達の影響

言語論の歴史を振り返ると
中世編

I. 文法（品詞論、統語論、語用論）：ポールロワイヤル
II. 観念の表現：ロック
III. 意味の素性への分解：コンディヤック

- 構造と意味→現代的な問題は出揃っている

I. 印刷技術のための統一された言語の構築：キャクストン

- 印刷という実用的問題から言語を制御：グーテンベルグの印刷の発明は、多くの哲学者や言語学者が束になってもかなわないほどの影響を言語研究に与えた

言語論の歴史を振り返ると
近世編

I. 真の言語を求めて→
   I. 古代の言語だがギリシア語よりも整ったサンスクリット語（屈折型言語）→屈折型言語の生産性の高さ
   II. インドヨーロッパ祖語：フンボルト
   III. ダーウィニズムが言語の系統を辿ることを刺激した

- そして革命が

ソシュール

- 思想は星雲のようなもので、その中で必然的に区切られているものは何もない
- 言語が現れる以前は何一つ判別できるものはない
  - 言語の恣意性
  - 言語の共時態を対象にした研究
  - 言語を遡るような研究をしても所詮は後知恵
  - 言語の構造を明らかにすること
  - 語が世界とどのように関係しているのという問題は言語研究の本質ではないと論破した

自然言語に関する科学ーソシュールの革命

- Saussure：ソシュール
  - 共時的（つまり同時刻の）言語システムの総体を　langue
  - 実際に使用された言語の現れ　parole
- langue　の構造を対象する科学としての言語学　linguistics
- 現代の計算機のよる自然言語処理は、ソシュールの延長線上にある部分が多いが、langueを基礎にしつつparoleにも対象を拡大

自然言語に関する科学とは

- ソシュール以前は、自然界の諸物に言語で名前をつけると思っていた。（言語命名説）
- ソシュールは混沌とした自然界は言語を用いて初めていろいろなモノに分節できる（つまり別のモノとして認識できる）と考えた。（従来から180度転換）
- つまり言語の自立性が主張された。よって、自然界から独立して言語だけを対象に科学できるようになった。

- 言語の自立性→
  - signifant　←　signe　→　signifie
  - 発音、つづり　記号　概念（対象物）
  - signifant,signifie とも言語に内在する。外界のものではない＝言語の自立性
- 恣意性
  - 記号、つづり、発音、概念のつながり方は恣意的に決まる。（枠組みは分かるが、なぜ？）

C.S.Pirce

- ソシュールのsignifiant vs signifie、および恣意性に対してパースは人間の認知過程まで射程に入れた。
  - コンテクストに言語を位置づける「解釈」を導入
  - 以下の3項組みによる
  - 左から右に進む（抽象化）

icon	index	symbol
abduction	induction	deduction
名辞	命題	論証
ソシュールは言語の独立性からここを対象外とした	signifant	signifie

- 演繹推論
  - 演繹規則だけで推論。公理系が与えられれば、真の命題は既に確定している。
- 帰納推論
  - 多数の個別規則から一般規則を導く。
    - 人→死ぬ、星→死ぬ ➔ 全てモノ→死ぬ
- 仮説推論(abduction)
  - 規則と与えられた結果から実世界についての仮説を導く
    - Aは死ぬ 、人→死ぬ ➔ Aは人
    - 嘘っぽいが、蓋然的
  - 日常の推論、日常の言語、実世界の鏡としての言語
  - 言語と実世界の関係付けは依然として未解決。
    - ロボットなど実世界で活動経験を持つ人工知能から新たな知見が得られるか、どうか。

計算機で言語する　チョムスキー

- 共時的Langue を全て網羅することは不可能
- この不可能に挑戦するのが言語学者
  - 特定の現象に特化した研究。例えば、「は」と「が」の差異
  - 「ワインが好きだ」vs「ワインは好きだ」
- 言語学者は自分たちが見聞きした言語現象から推理するしかなかった。
  - ただし、言語学者が記憶し整理している文例の大きさは膨大なものである。

計算機で言語する　チョムスキー

- しかし、Chomsky ：チョムスキーは言語能力は遺伝子に組み込まれているという立場を採っている（生得的という）。したがって、自分の言語能力を使って langue の本質に迫れると考える。
  - 例：John kills him. (him != John)
  - 　　John kills himself.
- 当然の帰結として、扱う対象は無意識に行われる文法（Syntax）までで、意味論は研究対象にならない。

計算機で言語する歴史

- 1940年代の計算機誕生とともに言語を計算機で扱う研究は始まっていた。
  - IBMのLuhnが1950年代初頭に既に計算機で文書から抄録を抽出するシステムを提案していた。
- 機械翻訳を目指した研究が盛んになった。
  - 1960年代の ALPAC(Automatic Language Processing Advisory Committee)レポートで機械翻訳が不可能と断定されたが…..

認知革命

- 認知革命以前の問い：言語の科学は物理学のように演繹的に構成できるのか？(1950年代)
  - データのみから帰納する。直観を排除：構造主義
  - しかし、計算機パワーが貧弱だった計算のモデルを欠いた帰納だけでは大きな発展が難しかった。
- 1960年代：認知革命：人間の言語処理、情報処理についてのトップダウンモデル
  - チョムスキーの変形文法
  - ニューウェル、サイモンの問題解決：人工知能
  - 計算機の能力のそれなりの進歩による部分多し。

チューリングテスト

- チューリングテストをパスする自然言語処理機械を作るには？
- 大きな九九表
  - 文と意味の対応表、日本語文と英語文の対応表
  - これではごまかしみたい。本質が分かった気がしない。
  - 無限に多い場合を考慮すると対応表が爆発
- 無限の可能性に対応できる計算メカニズム
  - チョムスキー型、人工知能型アプローチ
  - 無限に多い文や文脈を計算モデルとして考えきれるのか？
  - 中川個人としては「分割と統治」の方法論しか思い浮かばない

Top down vs Bottom up
合理主義 vs 経験主義

- 陥りがちなことは、
- 現実のデータを見ない理論(TopDown)
- 理論的方向性のないデータ集積(BottomUp)
- 機械翻訳の研究の歴史を例に T vs B の葛藤の様相を示そう。

Bottom Up 旧世代：構造主義

- 思弁的だった言語学を科学にしようとした試み
- 収集した言語データを主観を排して？？観察し、言語の本質的要素を明らかにする。
- 動詞の接尾辞「て」vs「で」
  - 同じ「て」だが、鼻音の動詞「死んで」の後では「で」になる。
  - 鼻音 vs　非鼻音　という相補分布でなければいけない。
  - 最小対(minimal pair)の考え方：
- しかし、「死んで」と「生きて」を同じカテゴリーだと見るのは全く主観を排して議論できるのだろうか。

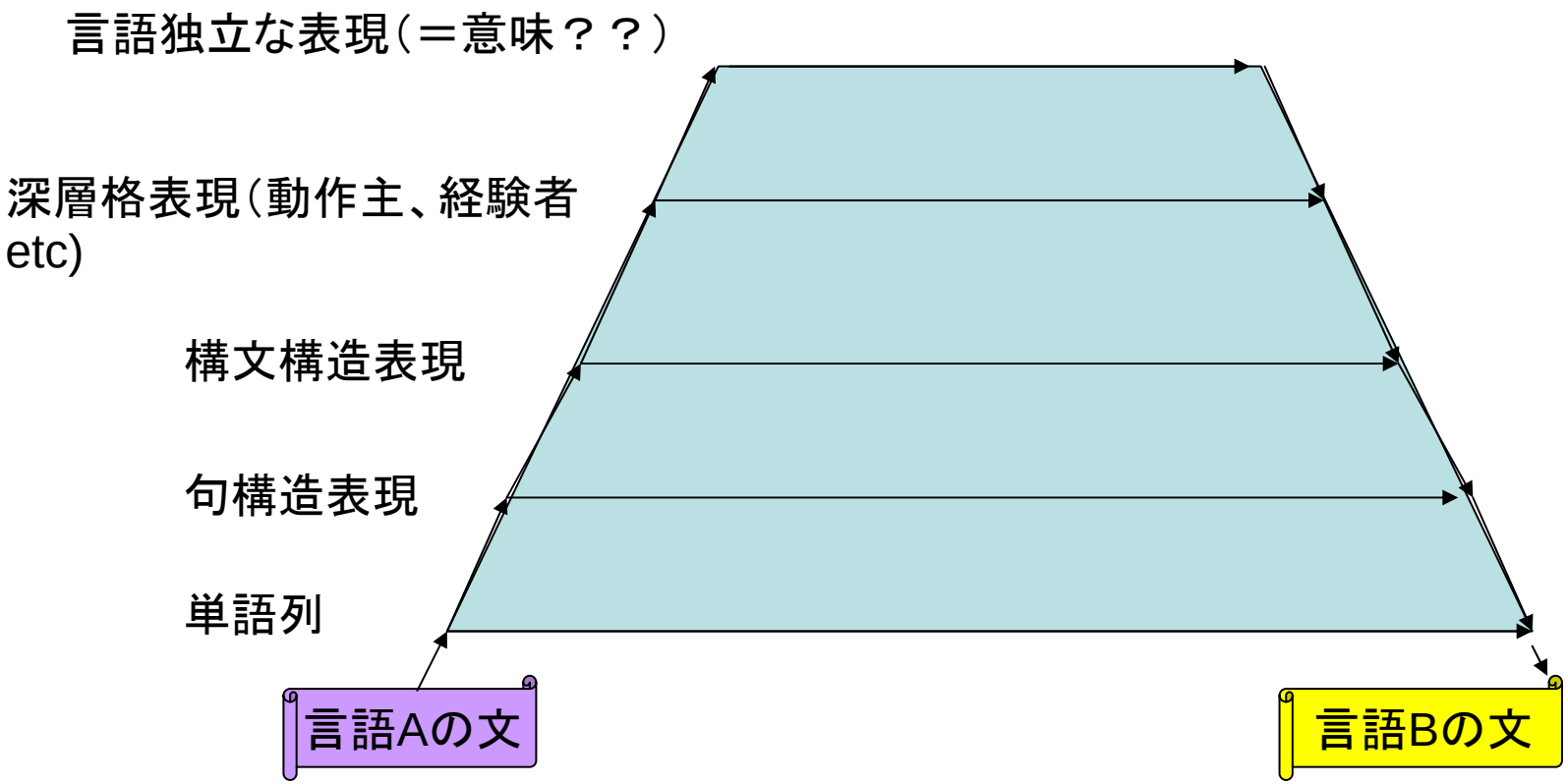
合理主義

- 出発点：言語から独立した計算のモデルを想定
- できるだけ単純なモデルが見通しがよい。
- 言語を実世界から切り離したソシュール的アイデア
- 最初はパフォーマンスが悪いが、いずれはBottomUpシステムを上回る。BTは現実のデータしか見ないから、予測能力が低いのだ。
  - しかし、最初のモデルが外れだったら？

- チョムスキーの個別言語に依存しない言語理論（普遍文法）に依拠
- 言語だけを相手にしたとき、自立した言語のモデルは構文論が最適

移行派原理主義：transfer fundamentalist

- 下図のどこかのレベルで言語Aから言語Bに移行する。
- 移行するレベルにおいては、言語Aと言語Bの表現の間で変換対応表を作れる（という信念）
  - たとえ対応表が膨大でも



移行派原理主義の問題点

- レベルが上がるにつれて構造が大きくなる。それでも言語AからBへ移行できるのは、
- 部分の意味は一度決まると、それを組み合わせることで全体の意味が決まるという構成性原理を前提にしてるからなのだが……
- 言語A,B間で単語の対応は一意的でない。
  - 湯、水　→ water
- 一方の言語にしか存在しない文法的性質や機能語あり
  - 冠詞、名詞の性
  - それでも複雑な変換表を作ればなんとかごまかせるかも

移行派原理主義の問題点

- 最も深刻なのは
- 意味の文脈依存性
  - 名詞の単数、複数の区別のない言語Aからある言語Bへ変換するには、文脈情報が必要。しかも文脈の数は無限。
  - デフォールトを単数に変換し、文脈で証拠が出れば複数と変換。
  - 「けっこうです」➔”thank you” or “no thank you”
    - デフォールトでは解けない！？

記号について
-- 少し視野を広げ人工知能の視点から--

- 記号と公理系から閉じた知識体系を作る(前記ヴィトゲンシュタイン)
  - 記号はそれ自体でひとつの存在。記号を用いた推論は、想定する集合上での操作として定義できる(外延的論理)
  - 80年代までの人口知能はこの路線だった。なにしろ、入出力が貧弱で計算機の外側の世界と通信できなかったから

- しかし、限定目的の貧弱なシステムしか作れなかった。（エキスパートシステム）
- 80年代後半から外界とのインタラクションが重視されるようになった。
  - ロボットにおける subsumption architecture
  - 分散知能
  - エージェント（これは現在ではソフトウェア工学）
- 文脈情報を考慮した記号処理への動き

文脈情報を考慮した記号処理への動き

- 記号は、
  - a. コアになる意味
  - b. 文脈に依存した、つまり言語使用における意味
- からなる。
- そこで、b.を考慮するために事例を大量に集めて事例ベース翻訳が考案された。
  - 翻訳事例
    - 「太郎は小説を読んだ」 vs “Taro read a novel”
    - には太郎＝人間、小説＝文字メディア、という文脈によって「読む」を規定する力あり。
  - しかし、それにしても個々の単語のコアな意味は予め与えないと動かない。

単語の意味

- 単語の意味を要素に分解して表現する方法（80年代）
  - Kill = cause (someone (alive → death))
- 何を基本要素におけば十分なのか？
  - 90年代以降の主流は
- その単語が使われた文脈に共起する単語で意味の曖昧さを解消する。
  - 大規模コーパス（20ヶ月分のNYタイムス）で、capital の資本、首都の意味の曖昧さ解消などが90%の精度でできた。
  - 未知語の翻訳も文脈に共起する単語の類似性を使って推定する方法が提案されている。

経験主義あるいはデータ主義

- 文脈あるいは言語使用における意味というデータ主導の方法をもっとラディカルにするのが経験主義
- IBMの統計的機械翻訳（90年代初頭）
- 人間でも気がつかないような英仏の言い回しの翻訳を純粋に機械的手法（統計的機械学習）で発見した。
  - EM, ビタビ探索など
  - 大量のメモリと高速な計算機
  - 大量の質のよい翻訳文の対(教師データ)
    - これがなかなか簡単に入手できない

計算機で言語する20世紀終盤

- 1970年代に計算機パワーの向上により機械翻訳は現実のものになった。
- 言語学の知識を用いたシステム
  - 言語学は、言語使用の広範な現象はカバーしていない。
  - 限定された現象の分析。例えば、「は」vs「が」
  - 1980年代になり計算機科学者たちが独自に文法を構築しはじめた。
- 正しくきれいな書き言葉の文法だけでは、実用性がない
  - 言語学の規則も現実の言語現象で正しい場合は60%？
  - 現実の言語現象はあまりに多様かつ広範

自然言語に関する科学とは

- 言語と実世界との関係はさておき、今できることは？
- 機械翻訳は、翻訳元、翻訳先とも言語だから、言語の中だけで完結できる。現在の機械翻訳はそのような構造。
- 文書分類、検索、要約、言い換えなども言語の中だけで完結型。
  - 画像とテキストが絡んだ場合はたちどころに困難が現れる。
  - 言語の中だけで閉じた言語学だけでは、自然界や人間界に影響を与える計算機システムは作れないこともある。
    - 例えば、計算機と人間のインタフェースを言語で行おうとすると、困難を生ずる。
    - ロボットに「これをあのごみ箱に捨てて」と命令すると、それを解釈するには外界のモデルが必要

計算機で言語する1990年代以降

- 自分の直感に頼っているのは本当の科学か？
  - 言語学の規則も現実の言語現象で正しい場合は60%？
  - 現実の言語現象はあまりに多様かつ広範
- 現実の言語データを大量に収集して分析したり文法を網羅的かつ機械的に獲得できないか
  - 統計的自然言語処理（90年代以降の主流）

計算機で言語する

- 音声認識
  - 書き言葉だけではなく話し言葉文法の必要性
- 大規模コーパスが出現した
  - 計算機処理可能な大量の電子テキスト（ギガバイト級）
    ＝　コーパス
  - 新聞記事10年分が計算機で処理できるようになって、いろいろな問題が見えてきた。
- ここで問題が生ずる
- 果たして広範な言語現象を文法として記述しきれるのか？
- 十分な言語データが入手できるのか？

計算機で言語する　現代の問題

- ここで問題が生ずる
  - 果たして広範な言語現象を文法として記述しきれるのか？
  - 十分な言語データが入手できるのか？

- 狙いをつけた言語現象に対応するデータが見つからないことが多い。
  - data sparseness の問題。
  - 例：全ての2単語の連続する確率を求めようとしても、多くの2単語連続は言語データに出現しない。
  - 統計的な小標本理論により、予測精度を向上させるという方向
  - 言語学者の知見も参考にできればする。

- 現実には、質の悪い翻訳対データでなんとかしないと
  - 対訳でない場合。同じ内容について、あるいは同じトピックについての述べている2言語コーパス
  - 基本語彙の辞書くらいはある
  - 計算機は早いし、記憶容量も大きいとは言え
  - 機械学習パラダイムもなんとなく出尽くした？？
  - 人間との共同作業？？