私が期待する今後の大学における言語処理研究

関根聡
楽天技術研究所　・　ニューヨーク大学
sekine@cs.nyu.edu

1. はじめに

これまで、大学、ベンチャー企業、中堅企業の研究所の経験をしてきた私の目から、今後の大学における言語処理研究に対する非常に個人的な期待を述べたいと思う。言語処理は言うまでもなく工学的な分野であり、いかにユーザーにとって便利で役に立つものが作れるかが評価基準である。それは、エンドユーザーへの便利さの提供や、それによる金銭的な成功のみを意味するわけではなく、多くの人（この「人」にはアプリケーションにより近いシステムの開発者や他の研究者も含まれる）に使われる技術、ツール、知識を作りあげることもその評価基準に含まれる。このような評価基準を共有しながらも、大学、ベンチャー企業、企業の研究所では環境や目的が異なっており、それぞれに最適な言語処理研究の方向性があるように思う[1]。まずは、大学、企業の研究所の環境や目的についてまとめ、そこから導かれる私の大学における言語処理研究への期待について述べたいと思う。

2. 環境の違い

表１に、大学、ベンチャー企業、中・大企業の環境の違いについて一般論をまとめてみた。[2]

	資金	資金の流動性	主目的	他の目的	人・データ
大学	そこそこ	既に成功した人に多く投資される	新しい価値の創造	人の育成[3]	学生はいる。データはあまりない
ベンチャー企業	ない	短期的に儲かることを説得できれば投資されることもある	儲かる仕組みの構築	わくわく感の創造[4]	人もデータもないことが多い
中・大企業	結構ある場合が多い	短期的に儲かることを説得できれば大いに投資される	儲かる仕組みの維持・発展	社会へ貢献	定常的には人は少ない。データはある

表１．大学、ベンチャー企業、中･大企業における言語処理研究にまつわる環境の違い

[1] 2011 年 3 月 11 日に行われた言語処理学会年次大会ワークショップ「自然言語処理における企業と大学と学生の関係」では、情報通信研究機構の鳥澤健太郎氏は上記の意見とは真っ向から反対する意見を述べている。ちなみにこの日は東日本大震災のあった日で、岩田氏の招待講演の最中に長い揺れを感じたことが記憶に残っている。震災の被害にあわれた方々に再度追悼の言葉を述べたい。また、この脚注に挙げた方々以外にもそれぞれの経験に基づいた非常に参考になる話が含まれているので、興味ある方には引用にある URL を覗いてみて欲しい。

[2] 同ワークショップで東京大学（当時）の荒牧英治氏、国立情報学研究所の宮尾祐介氏、株式会社はてなの田中慎司氏は似たような分析を述べている。

[3] 同ワークショップで東北大学の乾健太郎氏、グーグル株式会社の賀沢秀人氏、サイボウズ・ラボ株式会社の竹迫良範氏、株式会社ミクシィの木村俊也氏、楽天技術研究所の萩原正人氏が非常に参考になる意見を述べている。

[4] 同ワークショップで株式会社アルベルトの上村崇氏、ロックオンの岩田進氏がベンチャー企業に関して非常に参考になる意見を述べている。

表の内容については、異論もあるかもしれないが、詳細に説明しないでも大体の内容は理解してもらえると思う。大学は、お金はそこそこあり、人がいて、データはない。目的は新しい価値を作ることである。ベンチャー企業は、お金も人もデータもない。目的は儲かる仕組みを作ることである。中・大企業の目的は既に儲かっている仕組みを維持・発展させることであり、定常的には人は少なく、データはある。短期的に儲かることを説得できれば、人もお金も付くことが多くある。

3. 大学への期待の背景

上記の環境の違いを元に、大学での言語処理研究の方向性を議論したい。まず、大学で既に成功した人に資金が多く流れ、より面白いことや新しい価値の創造ができる可能性が広がることが言える。まずは、多くの言語処理研究者が大学で成功することが、私の大学での研究への期待の第一歩である。ところで、「大学での成功」の定義が必要である。様々な定義があるとは思うが、これまでに「成功した」と言われる人々の業績を見ると、１）新しい分野を開拓した。２）世界の一流の研究者からも素晴らしい研究だと認められた。３）他の研究分野の人からも話を聞きたいと言われるようなその分野の代表になるということが挙げられよう。大きな研究資金の調達はその結果でしかない。まだ資金を調達することが必要ない若い世代の人の場合には次のような一般的な話を紹介したい。まずは、若い人は研究室で頭角を見せ小さな成功を積み重ねて行くことが重要である。研究室の人に認められなければ、世界の人に認められるわけがない。そのような成功を積み重ねると、だんだん大きなプロジェクトがやりたくなってくるし、そのようなものを任されるようになってくる。そして資金を持つ教授などのグループのメジャーなメンバーになり、その後徐々に自ら資金を獲得できるようになって行く。小さな雪だるまを転がして大きくして行くように、少しずつ大きな成功を繰り返し実現することがとても重要であろう。

今度は、言語処理研究の内容について大学に期待したいことをまとめてみたい。上記に個人における成功の話を書いたが、研究も同じようなことが言える。研究の進展は、過去の研究を雪だるまの核としてその周りに雪を張り付け大きな雪だるまにしている過程に似ている。例えば、これまでの形態素解析の方法論の発展を見てみると、最長一致法、コスト最小法、機械学習によるコストの計算などが前の方法に皮をかぶせる形で進んできた。ただ、大きな目で見ると、現状における言語処理の一つ大きな問題点として、文字として表されたシンボルの操作（manipulation）をしている限りでは到達できない領域があるように思える。意味の世界である。形態素解析、構文解析、固有表現抽出などはシンボル操作の範囲で 90%以上の精度を得られる課題だったかもしれない。しかしながら、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、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。この問題は 2008 年に行われた”NSF Symposium on Semantic Knowledge Discovery, Organization and Use”で筆者の発表に含まれた図 1 で説明したい。この図では、1960 年代には対話処理が、70 年代には情報検索が、80 年代には情報抽出と自動要約が、00 年代には質問応答が新しい課題として提案されて研究が盛んになったが、10 年くらいの時間がたつと 60%程度の精度まで到達し、それ以上の進展を見せないまま研究が下火になり、次の課題に移っていったという状況を表している。最近では 2000 年頃に質問応答が新しい課題として情報検索を引き継ぎ形で提案されたが、IBM のワトソ

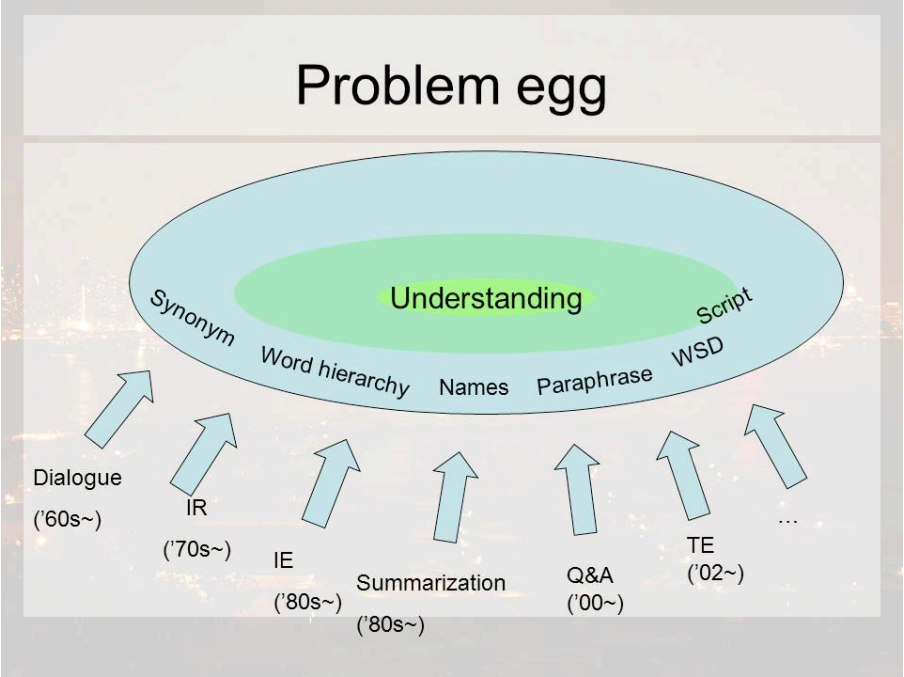

図 1.　Problem Egg

ンのある程度の成功を見て下火になっている状況だと分析している。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、それを壊せないで衰退しているという状況なのではないかと考えている。そして、私が推測するに、その卵はすべてが「意味」という同じ卵なのではないかと考えている。つまり、シンボルとして分析対象としている文章には現れない情報。人間が文章を理解するときにすでに頭の中に入っている世界知識。このようなものが文章を理解するためには必要となっているはずである。この問題は、もちろん、言語処理研究の研究対象にはなってきているが、捉えどころがなく形式化しにくい問題ではある。現状、挙がっている課題としては、同義語、意味の階層、名前、パラフレーズ、語義の曖昧さ、スクリプトなどが挙げられるが、これだけではないであろう。このような問題の解決に取り組んでいくことを、私は今後の大学における言語処理研究として望みたい。

4.　大学への期待：「意味の研究」

意味の研究は難しい。すぐに人の役に立つものができるかどうか分からない。長期的な問題である。まだ解決案があるわけではなく、未知な分野であり批判的思考力の養成に適している。まだ解決案があるわけではなく、それができた暁には新しい価値の創造が絶対に達成できる。今できていないということは、そのような新しい価値がそこに眠っていることに他ならないからである。とりあえずの意味の研究では今大学が持っている以上の大きなデータの必要性がそんなに重要ではないように思われる。大学では儲かる仕組みの構築、維持、発展が必要なわけではない。大学はわくわく感を直接作り上げて行く必要はない。表 1 を見直して欲しい。これ以上、大学がやるのに適した言語処理の課題があるであろうか？意味の研究を大学がやることは「私が期待する課題」というよりも、合理的に考えた場合に大学の価値観や存在意義を最大化する最適な課題

のように思えてならない。問題があるとすれば、優秀な学生の確保であるかもしれないが、「コンピューターが人間の言葉を理解する」という壮大な夢に向かった大事な一歩であることを強調すれば、浮き足だっていないしっかりとした思考能力のある優秀な学生を集めることは可能なのではないだろうか？

最後に、「どのように進めるべきか」という疑問に私なりの答えてみたい。私は、今は 100%の精度で処理することができない技術の分析から始めるのが真っ当だと考えている。なぜ、対話システム、情報検索、情報抽出、自動要約、質問応答は 60%の精度しかえられないのだろうか[5]？それぞれの分析対象において、どのような知識や技術があれば精度が向上するのであろうか？この地道な分析と努力が、言語処理の新しい道筋を示してくれるような気がしてならない。大学の研究として機械学習による精度のチューニングや大量のデータの簡単な再フォーマット化による見栄えのよさだけで満足するのはやめようではないか？その次が重要であり、皆で Problem Egg の殻を破る方向に向かっていこうではないか？

参考文献

言語処理学会第 17 回年次大会ワークショップ「自然言語処理における企業と大学と学生の関係」豊橋科学技術大学 2011. http://nlp.cs.nyu.edu/gengo2011ws

NSF Symposium: Semantic Knowledge Discovery, Organization and Use, November 14-15, 2008. http://nlp.cs.nyu.edu/sk-symposium

[5] 例えば、IBM のワトソンが Jeopardy の最後の問題「その都市の最大の空港は第二次世界大戦の英雄の名前を取り、2 番目に大きな空港は第二次世界大戦の戦いの名前から取っている米国の都市（US City）はどこか？」という質問に対して「トロント」と答えた。この間違えに対して、「トロントという名前の都市は米国にいくつもある」「カナダの大都市トロントの一番大きな空港は第一次世界大戦の英雄の名前から取られている」「US の同義語である America という単語はカナダも含む意味でも使われる」という説明がされている。正しい正解である Chicago が 2 番目の候補になっていたようであるが、ワトソンが取った質問応答の方法論を詳細に分析し、このような間違えがでないための知識や技術を構築して行くことが重要である。もちろん、大学の研究ではワトソンのシステムから始めることはできないと思われるが、すべての 100%の精度に達していないシステムにおいて、このような分析が出発点であろう。