

研究の場としての評価型ワークショップになるために

関根聡*1 影浦峽*2 奥村学*3 乾健太郎*4

*1ニューヨーク大学

*2情報学研究所

*3東京工業大学

*4奈良先端科学技術大学

概要

1998年に始まった2つの評価型ワークショップ NTCIR-1とIREXの開催から今年で7年になる。この間に、データの蓄積や研究者の裾野の拡大という目的は達成されてきている。しかしながら、現在の参加者の状況を見ると、順位向上を目的とし、従来の技術の洗練を行うだけの手法が多く見られ、新しい技術の開発や新しい研究の芽の発見が、その全体的なコストに対して少なく感じられる。本論文では、その状況を鑑み、評価型プロジェクトのあり方や目標の設定の方法などについて提案を行う。それは、参加者からの活発な議論を通じたボトムアップなテーマや問題の設定、現実的な応用を視野に入れた形での運営、結果の詳細な分析を通じた技術の洗練であり、それらを総合した所の「考える場」としての評価型プロジェクトというあり方である。

1. はじめに

「評価型ワークショップを考える」というワークショップに対して、評価型ワークショップであるIREX、TMREC、TSCを主催していた3名と評価に参加している1名が、現在の日本での評価型ワークショップに対しての意見を交換し、その見解がほぼ一致したので、本論文ではこの4名の「考え」を述べたいと思う。

まず、ワークショップを主催することは、研究的な側面だけでなく多くの苦勞がある。NTCIRを5回になるまで牽引してきたNIIの神門さんを始めNTCIRの関係者、その他、参加者を集め、評価型ワークショップを実施してこられた方々にも同様に敬意を表したい。

さて、本論文では現在行われている評価型ワークショップについて意見を述べるのであるが、特定の個人やタスクを批判しているものでは決してない。ここで述べたようなことがすでに考えられている場合もあるし、考えているけど実行できていないだけという場合もあるだろう。あくまで一般論として受け取っていただきたい。研究を実施するにはいくつもの方法論があり、評価型ワークショップはその一つであるにすぎな

い。そして、現状の評価型ワークショップについての問題点を明らかにし、改善していく道筋を提案することが本論文の目的である。

2. 目的はどれだけ達成されたか

評価型プロジェクトが日本で始まった当初、どのようなことが目的でこの試みが始まったか、NTCIR-1の予稿集とIREXの予稿集を紐解いてみた(NTCIR-1 1999)(IREX 1999)。それぞれNTCIRの目的は4つ、IREXは5つに箇条書きで述べられているが、総合するとほぼ以下の3点にまとめられる。それぞれのポイントについて現状を分析してみる。

データコレクションの作成

少なくとも、情報検索、要約、質問応答、用語抽出、固有表現抽出などの分野で、評価型ワークショップで作成されたデータは、それ以前には存在していない規模や内容のものであり、現在は各分野の研究で広く使用されている。データコレクションを作成し、それを幅広い人に使ってもらい、研究の推進に役立てたいという当初の目的は達成されていると考える。しかし、反面、与えられたデータを使うだけで、データの意義や問題を直視し考えるきっかけが失われかねないという危惧がある。

研究者の交流、拡大

評価型ワークショップが存在していない状態では研究推進が困難であった若い学生や、言語横断タスクなどへの参加による海外の参加者が、幅広く評価型ワークショップに集い、研究者の交流が行われている状態は非常に望ましいものであると思う。この点も目的は達成されていると考えられる。しかし、数多くの参加者が集まり、評価が行われたということ自体がワークショップの成功の基準ではない。参加者間での刺激が新しい技術の開発や新しい研究の芽の発見につながるということが重要である。その点が十分達成されているかどうか判断するのは難しい。またその目的のために

は「自分の頭で考える」積極的な参加者が増え、その間での交流こそが重要なのだと考える。

研究の推進

評価型ワークショップにおける参加者の最終的な目標は、新しい技術の開発や新しい研究の芽を見つけ、自然言語に関する技術の発展を実現することのはずである。しかし、筆者らの知る限り、現在の評価型ワークショップでは、既存の技術をチューニングさせたり、知識の量を労力によって増やしたり、非常に優れているといわれる他の人の技術を持ってきたりするだけでシステムを作成し、順位を競っていたり、参加することにのみに意義を見出している参加者が少なからずいるように思う。また、主催者側にも、主催すること自身が目的となってしまっていて、自分の研究目的なり問題意識を基に、その解決の手段としてワークショップを主催しているという意識が薄い場合もあるように思える。参加者も主催者も研究の推進が最終目的であり、そのためのワークショップへの参加であり、そのための主催であるという認識を常に持ち続けることが必要であると思う。次の章で、具体的な問題点について述べる。

3. 何が問題か

研究について

いくつかの異なる分野に横断的に関わっていると面白いことに気づく。大きな傾向として、一見単純で自明と思われている現象や事象に複雑さを見いだすことが研究において最も重要な点であると見なす研究者と、(複雑な)問題に(単純な)解決を見いだすことが重要であると見なす研究者とがいることである。さらに言えば、後者のタイプの研究者は前者のタイプそして自らの立場にどちらかというとならざるを得ない。前者のタイプはどちらかというとならざるを得ない。前者のタイプはどちらかというとならざるを得ない。

むろんこれは、科学革命と通常科学、問題発見/Aジェンダ設定と問題解決、俗流科学哲学的な帰納・演繹のサイクルといった概念設定に近づいたり重なったりする、繰り返し指摘されてきた凡庸なことである。

それにも関わらず、まさに言語の領域が極めて複雑かつ膨大であるとの抽象的自覚ゆえにこそ、解決不可能ではないものとして抽象化され部分化され限定されて取り出された一定の言語研究アジェンダが、自明の出発点として研究の全体を覆い尽くすに近しい状況に到るならば、そしてそれゆえに当初存在した言語の領域の複雑性と規模が忘却される結果を導くならば、科学

論の平凡な指摘を繰り返しておくことは無意味ではなからう。

この観点から見ると、残念ながら、評価型ワークショップは問題設定の自明性と全般性を必ずしも健全とは言えないかたちで促す結果となっているとの危惧は拭えない。もちろん、オーガナイザや参加者の多くは「その程度のことは知っている、しかし・・・」と言うだろう。けれども、ある種の罪を犯す人々の多くは反省的認識が発動する限りではその行為がいけないことは知っている。にもかかわらず直接的現場では犯罪に到る。したがって、個々の反省的認識のレベルで、これまで繰り返されてきた科学や研究の議論を踏まえることだけでは、恐らく十分ではない。

そこまで抽象的なレベルに議論をもって行かなくても、とりわけ 1990 年代以来自然言語処理研究の主流となった統計的・機械学習的手法について、そうした領域の草分け的存在である赤池弘次の次のような指摘を改めて思い起こしておくのが有用であろう。(赤池 2004)

「データによって、当面の意図に沿う適切な行動に導く知識を獲得するのが、統計的推論の目的である。これは、データの発生機構に関する仮説の提示とその検証の過程を通じて実現される。適切な仮説を提示できない人は如何に計算法をリファインしても目的を達成することができない。すなわち、形式的な統計手法の知識ではなく、対象についての理解、知識、そして経験等が決定的な役割を果たすということである。」

評価型ワークショップにおいては、しばしば問題の同一性が結果により維持されるため、類似した結果を導く仮説の多様性さらには仮説が位置する認識の枠組みの多様性が十分に論ぜられない結果となる。とりわけそれは、若手研究者の育成といった関係で長期的に大きな問題となる恐れがある。

NLPの研究者として

機械学習の研究がこれほど発展を見せ、学習でできることに関する理解が深まってきた現在、NLP 研究としては、問題を切り取り、タスクを設計し、タグを設計する過程そのものがますます重要で本質になってきている。この過程をオーガナイザや他人に任せてしまうという姿勢は、NLP 研究者としては健全でない。機械学習の研究者ならわかるが、そうでない NLP の研究者の個々がこの過程を放棄してしまえば、NLP 研究は本当に ML 業界に食われてしまう。従来型の評価型

ワークショップでは、下手をすると、この傾向を助長することにもなりかねない。もしかしたら、健全な研究者の養成を妨げる恐れもある。

とは言え、問題の共有・データの共有は技術の発展に欠かせない。これは、評価型ワークショップのジレンマと言うより、現在の NLP 研究そのものが抱えるジレンマとも言えるかもしれない。バランスが重要と言うのは簡単だが、その実現はそれほど簡単でない。

評価型ワークショップのような場で参加者が問題設定・タグ設計の過程に参加するのが一つの方向であるように思える。しかし、問題設定・タグ設計そのものが研究の重要な柱の一つであるとすれば、その「研究成果」を参加者共有の「研究業績」にするような形にしていくなかには何らかの工夫が必要である。

オーガナイザと参加者の関係

自動要約の評価型ワークショップ TSC(Text Summarization Challenge)は、過去 3 回にわたり、NTCIR の傘下で行なわれてきた。NTCIR での 3 回の経験から、現在の評価型ワークショップが、「ボランティアのオーガナイザ」と、残る「お客さんの参加者」の構図になっており、参加者は「受身」な存在で、オーガナイザに負担が集中するなど、弊害が大きいと考えるに至っている。

この原因の一端は、より多くの参加者に参加してもらうためという理由で、いくつもの配慮をオーガナイザ側が提供することにあるのではないだろうか。このことで、参加者はますますお客さんとしての居心地の良さを強めていっているのではないだろうか。その典型が評価費用の負担である。参加者が自分のシステムの評価分の費用を自己負担するなら、少しは考え方も変わるのではないだろうか。

共同で何かをしようというとき、本来ならそのメンバーは、原則「give and take」の関係にあるはずであり、それが理想に思われる。そういう意味で、評価型ワークショップにおいても、参加者は全員何らかの意味での(参加、結果提出ということ以外の)貢献が必要であり、それを求めていく運営が望ましいように思われる。したがって、そもそも、オーガナイザ、参加者という区分が存在していること自体に問題があり、全員でタスクを設計し(全員がオーガナイザで)、そのタスクに対して全員がシステムを持ち寄り(全員が参加者)というのがあるべき姿なのだろう。

夢の実現手段としてのワークショップ

評価型ワークショップが健全に運営されていくため

にはワークショップへの参加者が研究上の夢を持ち、その夢を実現する手段としてワークショップを位置づける必要がある。特にこの夢はオーガナイザと呼ばれる人達に欠かすことのできないものである。既にやられた課題を無為に真似、それを実施するだけでは研究の大きな進展は望めない。ワークショップの運営の目的は運営ではないはずである。極端なことを言うと、データ作成と課題の認知ができたならば、運営のための運営は止め、新しい研究を始めるのが健全な方法ではないかという考え方もある。

ワークショップの参加者が研究上の夢を課題に落とす作業としての評価型ワークショップの例を挙げる。米国の要約の評価型ワークショップである DUC では、ACL 併設ワークショップや独自のミーティングを開き、将来の要約技術の姿という議論が頻繁にされている。Dr. Karen Spark Johns や Dr. Donna Harman といった大御所の先生らが夢を語り、それを実現していくための課題のあり方についての議論を、実際のデータにまみれて行っている。情報抽出、照応解析の評価型ワークショップである ACE でも、Dr. Ralph Grishman から大御所の先生らが実際に自分の頭と手を使い、地道にサンプルデータを作成し、次なる課題の検討を行っている。ここでも、最終目的は自分の夢の実現であり、その道のりとしての評価型ワークショップという位置づけである。

4. 具体的な提案

最後に、現状の評価型ワークショップを改良していく方向での具体的な提案を行いたい。これらのすべてが、確固とした主張の基での提案というわけではなく、評価型ワークショップに興味を持つ人が議論を始めるためのきっかけとなればという意味での提案である。

問題、テーマのボトムアップな設定

既に述べたように、自然言語の研究とは、評価型ワークショップに参加していい得点を取るだけのことでなく、極端に言えば、適切な課題を設計することであるといえる。したがって、評価型ワークショップには自然言語に対する夢や自分の考えが必要であり、その夢をぶつけ合い、その実現への道のりを設計する場が必要である。例えば、同じような夢を持つものが集まり「ロードマップワーキンググループ」といった場で次の課題を議論するのがひとつであると思う。ワーキンググループに参加者が積極的に集い、新しい課題への議論が活発に行われる状態が理想的である。参加

者は何も多数である必要はない。同じ課題に多くの積極的な参加者が集まるという状態は、もしかしたら逆に危惧されるべき状態であるかもしれない。とにかく、課題を設計するワーキンググループもなく、オーガナイザと呼ばれる人達がほぼ絶対的に課題を決めてしまう評価型ワークショップは非常に危険であり、その課題によっては単に「こなす」だけのワークショップでしかないように思われる。

しかし、現実的問題として、ロードマップワーキンググループを作ってそこで議論するだけでは、やはり若い研究者の参加に障壁を作りかねず、例えば、オープンな課題の設定のためだけのワークショップがあってもいいように思う。そこでは、課題やデータのデザインを議論したり、評価型ワークショップで作ったデータコレクションの性質、問題点などを分析し、新しいデザインのための予備調査をしたりする。そういった論文を公募して、予稿集をちゃんと出すことによって、若い研究者も主体的に議論に参加できる設定を整えることを目指す。

応用を考えた運営

自然言語の研究は、人間に役に立つシステムを作成することが究極の目的であり、すべての課題はなんらかの応用につながっていると考えられる。特に、現在、評価型ワークショップとして課題になっているようなものは、非常に応用に近い位置にあり、それがどのようなユーザーにどのように使われるのかということを中心に描いて実現することが必要である。実際のユーザーを招聘し、そのユーザーの満足度を評価の基準を作成する際の参考にするのが本筋であるように思う。

参加者はソースコードを公開する

現実的な問題として、著作権や企業秘密などの問題はありますが、目的の一つが研究者の交流であることを考えると、究極的には、ワークショップに参加したグループは開発したプログラムやデータを公開することが望ましい。例えば、IREX の発表ワークショップではシステム発表以外に、評価結果の分析を行った。TMREC でも、横断的な結果の分析にかなりの労力が払われている。しかしながら、どうしても参加者のシステムの詳細が分からないことには分析が進まないことが経験上わかった。研究者の交流とは、一緒に集い雑談レベルで研究の話をするのではなく、結果の分析や分析に基づいた改良や新しい問題の発見こそが交流の本当の意義であり、研究の推進につながると考える。プログラムやデータの公開を参加者に義務付けるなり、

公開する参加者に特典を与えるなどの処置があるといえるように思う。

ワークショップ毎の時間をあける

現在連続的に行われている NTCIR のワークショップの間隔は1年半である。しかし、経験上1年半では新しい試みに挑戦することができない。また、新しい課題を提案し議論していく期間としても十分でないと思われる。ある程度のデータができたならば、それと同じような課題を繰り返すことの意味はあまりないのではないだろうか。その課題で研究を続けて行きたい人は、そのデータを常に使い続けて性能の向上を見ていければいい。新しいデータは新しい問題がみつかり、その角度から課題やデータを見直す必要ができたときに必要になってくるものではないだろうか。

参加者の投票による優秀賞の設定

数値的な評価の弊害を除去するために、数値だけではない評価基準を作ることを提案する。既存の技術の洗練による最高点よりも、新しいアイデアや挑戦的な試みの方をより高く評価すべきであると考えためである。評価が終了したら参加者は他の全てのシステムについての論文を細かく読み、新しいアイデアや挑戦的な試みを行っているものに対して投票を行う。その合計点の多いものに優秀賞を与える仕組みである。これは、参加者にモチベーションを持ってもらう効果と共に、すべての参加者が他のシステムをしっかりと調べなければいけないことからくる、本当の交流の促進をも期待できる。

参考文献

- NTCIR-1 ワークショップ予稿集．学術情報センター（現 NII）1999．
- IREX ワークショップ予稿集 IREX 実行委員会 1999．
- 赤池弘次 2004. 「真理への近さを測る」山本光璋・鷹野致和編『ゆらぎの科学と技術-フラクチュオマティクス入門』仙台：東北大学出版会．p. 19.