

# Project Next NLP 概要 (2014/3 – 2015/2)

関根聡

ニューヨーク大学

sekine@cs.nyu.edu

## 1. はじめに

本稿では、2014年3月の言語処理学会第20回年次大会ワークショップ「自然言語処理の発展に向けた情報共有・討論」での議論を発端として始まった Project Next NLP の活動のうち2015年3月の言語処理学会第21回年次大会ワークショップ「エラー分析ワークショップ（兼 Project Next NLP 報告会）」（以降、「エラー分析 WS」と呼ぶ）の直前までの活動をまとめて報告する。なお、本プロジェクトは今後も継続予定であり、その様子は以下の HP に逐次アップデートしていく予定である。

Project Next NLP ホームページ：<https://sites.google.com/site/projectnextnlp/>

## 2. 背景

ホームページには本プロジェクトの背景、目的が以下のように述べられている。

科学技術が“trial and error”によって進展してきたことは間違いありません。この“trial and error”を科学技術の効率的な発展につなげるためには、偶然の産物を求めた無作為な試行の繰り返しではなく、「error を深く分析し次の有効な trial につなげる」プロセスが重要であると我々は考えています。

自然言語処理の技術は、様々な応用を目指して進んできました。1960年代には対話技術が、70年代には情報検索技術が、80年代には情報抽出、要約技術が、2000年頃からは質問応答が盛んに研究されてきましたが、それらすべての応用システムの精度は60%程度まで達成した後にそれ以上の精度向上が見られないまま、新たな応用を求めて移ろって来たという歴史があります。90年代から始まった大規模なコーパスを元にした経験的な手法により新たな展開が生まれ、統計手法や機械学習を巻き込み発展してきた事は事実ですが、未だ60%の精度の先に必要とされているものの正体がきちんと見えているとは言いがたい状況です。

そこで、この問題の正体を明らかにするために、様々な言語処理技術においてしっかりとしたエラー分析を行い、次に対処すべき課題を明確にする事は非常に重要だと思います。言語処理技術の各領域においてエラー分析はこれまでも個別の研究の一部として個々の研究者が行って来ていますが、そのほとんどは特定のタスクにおける特定のシステム・手法を対象とするもので、システム・手法に固有の問題が混在した形で課題が明らかにされてきただけでした。このプロジェクトでは、それぞれのタスクについて複数のシステムでエラー分析を行う事で、そのタスクにおいて共通に重要な問題点を明らかにしていきたいと考えます。また、1つのタスクだけではなく、多くのタスクのエラー分析を同時に行う事で自然言語処理全体における問題点の列挙ができるものと考えます。このような大規模なエラー解析は過去に経験がなく、どのような結果が得られるのかやってみないと分からない部分もありますが、今後の自然言語処理研究の方向性を決めるために重要なデータになる事は間違いのないと思われれます。一人でも多くの研究者に今回のワークショップに参加していただき、自然言語処理研究における次なる展開につなげていくことを期待しています。

つまり、複数の自然言語処理技術を「エラー分析」という一つの側面で切り取り、タスク横断的に自然言語処理という分野でのエラーの関係性、タスクの関係性、全体的な技術の方向性を探ることが本プロジェクトの目的である。

ここで、エラー分析に関連し、電動レジスター、点火装置、有鉛ガソリン、フロンを発明し、発明に関する名言・格言を多く残したことで有名なチェールズ・ケタリングの以下の言葉を引用したい。

999回失敗しても、1回うまくいけばいい。それが発明家だ。失敗は、うまくいくための練習だと考えている。(成功の99パーセントは以前の失敗の上に築かれる。)

失敗する事を恥ずべきではなく、その原因を明らかにする為にあらゆる失敗を分析すべきである事を良く教える必要がある。失敗するという事はこの世で重要な教育科目の一つなのだ。

*It doesn't matter if you try and try and try again, and fail. It does matter if you try and fail, and fail to try again.<sup>1</sup>*

上記の言葉は、エラー分析を地道に行う者にとってとても心強い。特に2番目の言葉は全体調整役としてこのプロジェクトを運営している筆者にとって常に心に留めておきたい言葉である。もちろん、自然言語処理の研究者なら失敗の重要性、必然性は十分に理解していると思う。人間の言葉の営みをコンピューター上で工学的に実現しようとしている自然言語処理技術の研究開発に取り組んでいる我々は、失敗(目標としているシステムが人が満足できる精度で動かないこと)を毎日経験している。満足できない精度しか出せない限り、そこにはなんらかのエラーの原因があり、それを分析し精度向上を目指すべきである。しかし、論文を提出すること、研究資金を得ること、評価型コンテストでいい成績をおさめることに注意を払い過ぎていると、時として失敗の重要性を忘れてしまっていることがないだろうか？また、失敗をするにも色々な失敗の方法がある。そのあたりに注意を払っているだろうか？自戒を込めて研究活動を省みると以下のような疑問が湧いてくる。

## 我々は上手に失敗しているのだろうか？

- エラーに慣れてしまっていないだろうか？
- 本質的ではない精度向上に満足してはいないだろうか？
- 何が本質か見失ってはいないだろうか？
- 難しい問題を避け、次なる目新しい課題に移ろい続けてはいないだろうか？

以下、個人的な経験を基に、上記の疑問について少し考えたい。話の内容から以下の文章が、個人的なエッセーのような形になっていることをご容赦願いたい。

---

<sup>1</sup> あまりに英語が美しいので引用した。以下のように日本語で表現できるであろうか。「挑戦して挑戦して挑戦して失敗することは問題ではない。問題なのは、失敗した後、再度挑戦することに失敗することだ。」

### エラーに慣れてしまっていないだろうか？

自然言語処理の解析精度は向上している。特定のドメインの形態素解析は98%という高い精度を得ているし、固有表現抽出も90%を超えている。一般的な人間のパフォーマンスと比較して、それを凌駕しているという報告も耳にする。しかしながら、常に人を満足させられるだけの精度は得ていない。まして、要素技術である照応解析や、応用技術である情報抽出、情報検索、翻訳の精度は60%というような数字で止まったままではないだろうか？例えば、そのような精度のシステムが「ないよりはまし」ということで商用システムに使われていることは我々の誇るべき成果ではあるし、そうやって工学的な応用を見つけ、人のためになるシステムを作ることが我々の目標のひとつではある。しかし、60%が65%になったこと、または大学の研究者として、「なんらかの応用に十分な精度に達した」ということで満足して良いものであるだろうか？それは通過点であり、そのような精度（裏返して言えばエラー率）に慣れてしまっていないだろうか？我々研究者は、人間の言語の営みをコンピューター上で工学的に実現することが目標であり、まだ十分ではない精度に満足していないか、常に自問し続ける必要があるように思う。

### 本質的ではない精度向上に満足してはいないだろうか？

「ベースラインと比較してF値がX%上がったため、提案手法の有効性が証明された。」という文言で締めくくられている論文を目にする。もちろん、自然言語処理を始めたばかりの学生の練習問題として、または、とりあえず、どのような素性が有効か調べるための実験という位置付けであれば問題ない。しかし、その比較したベースラインのシステムが提案手法に対して内容的な関係としてどのようなもので、そのシステムからの精度向上が本来その実験が目的としていた効果であるか分析され、そして、その後どんな種類のエラーが残っているのかがわからない限り、それは本質的な精度向上かどうか判断できない。もしかしたら、その精度向上は、提案手法が目的としていた精度向上ではなく、ベースラインが特異的に間違っていたシンプルなエラーを修正しただけかもしれない。もしかしたら、精度が向上してない結果の中に、提案手法が解決しようとしていた課題がより多く残っているかもしれない。実際、私自身、あるアイデアに基づいた情報抽出の実験をし、統計的有意差のある精度向上が得られた後に、実験結果の分析をやってみると、偶然性としか言いようのない理由によって精度が向上していたことが分かったことがある。論文を書く者として、このような分析を怠らずにやることは一つの義務であろう。ただ、ここで少し困るのは、8ページ程度の口頭発表論文であれば、査読者もそこまでの追及や著者と確認ができるわけではなく、統計的有意差があればよし、ということで論文が通ってしまうことである。このような疑問を持ってしまうような論文は、私の経験からすると実は往々にしてある（自戒を込めて）。我々は、論文を通すことを目標に研究をしているわけではなく、我々の言語活動の工学的な実現のために、今までよりも本質的に高い精度を出す手法を実現し、目標に到達しようとしているわけである。ある意味、論文とはその副産物に過ぎない。例えば、機械学習を使って、様々な素性による数多くの実験を特定のデータに対して行えば、適当な分散に従って、精度が向上する素性があるであろう。しかし、それがその結果のみによって、本質的に言語処理の精度向上に貢献するものであるとは言えない。

### 何が本質か見失ってはいないだろうか？

我々の多くは、人間の言葉の営みをコンピューター上で工学的に実現することを目標にしている。そうだとしたら、自然言語処理研究者にとって、機械学習はツールであって目的ではない。機械学習というのは非常に興味深く魅惑的な研究分野である。しかし、それ自身を研究するのならばそれなりの立場で行う必要がある。機械学習研究者にとって、その応用分野の一つとして自然言語処理があるということと、自然言語処理研究者にとって、機械学習をツールとして使うということは全く異なる。例えば、X という機械学習手法よりも Y という機械学習手法の方が、あるタスクで有効だったとしよう。しかし、その結果のみでは機械学習の研究者に対して有益かもしれない証拠を一つ提供しただけである。自然言語処理研究者としてなんらかの結論が得られたことにはならない。自然言語の本質を見失わないためには、データを見るのが重要である。

思い出話になるが、筆者が博士の学生としてニューヨーク大学のグリッシュマン教授から与えられた最初のタスクは、構文解析によって音声認識の精度を向上させるというものであった。その時にはすでに ACL の会長も経験していたグリッシュマン教授が最初のミーティングに持参してきた分厚い紙の束は、30 発話に対する音声認識結果の N ベスト結果（確か N=5）を自らの手で構文解析をした結果であった。その分析結果を使って、この研究がどのような意味と限界を持つかという説明からプロジェクトが始まった。まさにこのような分析からタスクや技術の本質を見つけていく必要があると思う。データを見る作業はとてつもなく大変である。まして、公開を目指してデータのアノテーションをする作業、アノテーションのための標準化、マニュアル化をする作業の大変さは十二分にわかっている。しかし、この作業を省いて自然言語処理の本質を見極めることは不可能である。すでに構築されたシステムについても同様であり、システムが出力した結果、その中でも特にエラーを自分の目と頭で分析し、その原因について深く考える作業こそが自然言語処理における研究の中核だと思っている。

### 難しい問題を避け、次なる目新しい課題に移ろい続けてはいないだろうか？

自然言語処理が対象にしたタスクとその課題を図 1 に提示した。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ぐ形で提案されたが、IBM のワトソンのある程度の成功を見て下火になっている状況だと認識している。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」というタスク共通の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状況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すでに頭の中に入っている世界知識や常識、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、もちろん自然言語処理研究の研究対象にはなっているが、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、固有名のグラウンディング、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。このような課題は奥行きがあり、興味深い問題であるが、別の論文での議論に譲りたい。

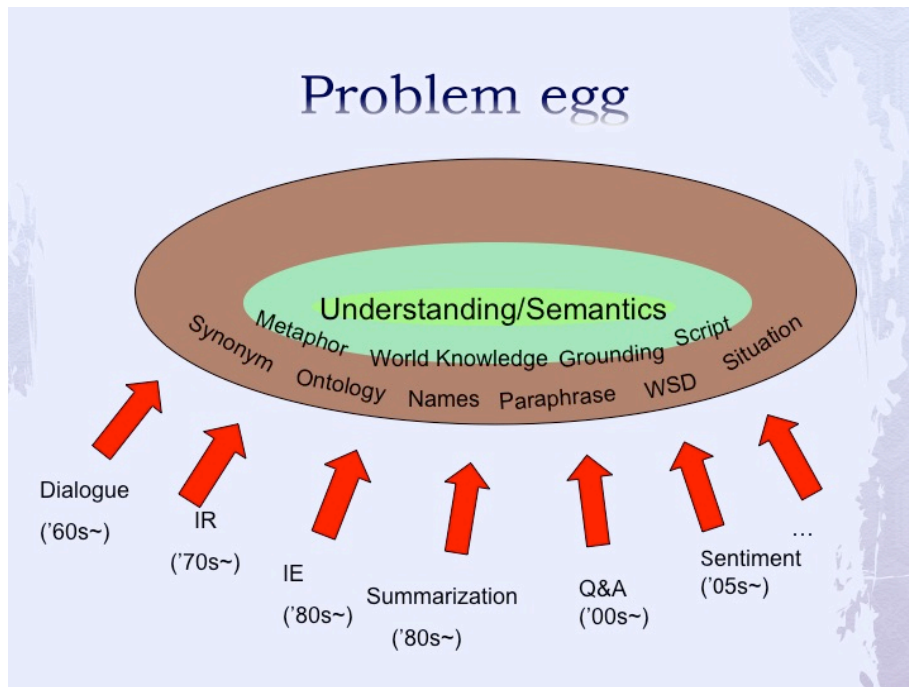


図1. 自然言語処理タスクの移り変わりと意味の課題

これに対し、「自然言語処理の発展に向けた情報共有・討論」のワークショップで乾氏が以下のように述べている。まさにその通りであり、乾氏のこの議論が本プロジェクトの一つの支柱である。

意味の問題に踏み込む研究は、関根氏も言うように、簡単ではない。たまたし、その最大の原因は、「意味」という同じ卵の中にあるというよりはむしろ、「60%」の先の課題かあまりに多様な個別的事項にあるのではないかという印象を持っている。照応解析や含意関係認識などで実際の解析誤りを見てみると、何か一つの根源的な問題がそこにあつて、それか解ければ一気に精度が上がるという構造になっている訳ではどうもなさそうである。「60%」の先には、オントロジカルな語彙知識や因果関係・スクリプトなどの世界知識が関わるすぐに思い浮かぶ類いの問題の他に、実に多様な問題が少しずつ顔を見せる。数量の計算の問題、時空間の推論の問題、領域ごとの慣習的スタイル、知覚にからむ問題、オノマトペ、メタファーなど、雑多な問題が少しずつ混じっていて、それらの誤りの蓄積で精度が伸びない。それぞれはおそらくかなり違う解法が必要である。しかし、どれか一つにアドレスして解法を作っても、それに関わる現象は元の照応解析や含意関係認識の評価用データにわずかしこ含まれないし、正解データにはノイズも混じっているので、全体の精度への寄与を定量的に測定するのは大抵の場合極めて困難である。したがって、なかなか元気が出ない。また、特定の問題だけを含むようなデータセットを作ろうとしても、不自然なサンプルになることが多く、今のところあまりうまくいっていない。

このように、「60%」の先の問題はかなり性質の違う雑多な課題の集まり・組み合わせになっており、それぞれを個別に解決しようとしても、その効果を測定する自然なデータセットを作るのが難しいのでなかなか研究が進まないという状況があるのではないか。最初に述べた自然言語処理の境界の融解がその傾向に拍車をかける可能性もある。言語の意味の問題に踏み込んでいくことは重要である。だからこそ、なおのこと、何がなぜ難しいのかをさらに深く理解していく努力とそれを踏まえた研究の方法論の議論が求められているように思われる。

### 3. 活動内容

本プロジェクトの活動方針は、自然言語処理でメジャーだと思われる複数のタスクに対して、同時にエラー分析を行う。それぞれのタスクにも複数の参加者を募り、複数のシステムのエラー分析を行い、それをまとめてもらっている。したがって、全体として得られたエラー文政は非常に多くのシステムのエラー分析担っているはずであり、自然言語処理技術の今後の方向性を見るための知見を蓄積している。設定されたタスクは18種類で、各タスクのリーダーは以下の通りである。(敬称略)

タスク	リーダー
<b>基礎技術</b>	
形態素解析	鍛冶伸裕 (東京大学)、森信介 (京都大学)
構文解析	河原大輔 (京都大学)
述語項構造解析	松林優一郎 (東北大学)
<b>要素技術</b>	
固有表現抽出	岩倉友哉 (富士通研究所)
照応解析	飯田龍 (NICT)
言い換え	藤田篤 (NICT)
語義曖昧性解消	新納浩幸 (茨城大学)
知識獲得	柴田知秀 (京都大学)
<b>情報アクセス応用</b>	
情報検索	難波英嗣 (広島市立大)
要約	高村大也 (東工大)、平尾努 (NTT)、西川仁 (NTT)
情報抽出	新里圭司 (楽天)
レビュー解析	藤井敦 (東工大)、乾孝司 (筑波大)
Web 応用	岡崎直観 (東北大学)、荒牧英治 (京都大学)
東ロボ	宮尾祐介 (NII)、横野光 (NII)、松崎拓也 (名古屋大学)
<b>翻訳、文作成支援、対話</b>	
翻訳	工藤拓 (グーグル)、グラム・ニュービッグ (奈良先端大)
日本語構成	山本和英 (長岡技科大)、鄭育昌 (富士通研)
英文校正	水本智也 (奈良先端大)
対話	東中竜一郎 (NTT)、船越孝太郎 (HRI)

リーダー選出の経緯は以下の通りである。まず2014年のWS後に筆者がタスクの候補ごとに連絡担当者をご指名し、担当者に各タスクにおけるエラー分析の方法論や可能性を考えていただいた。同時にそのタスクのグループを立ち上げることを考えていただき、リーダーを自薦、推薦していただいた。それ以外にも立候補や他薦などの形で就任していただいたケースもある。また、2つのタスクが1つにまとまったり、連絡担当者はいたが成立しなかったタスクもある。タスクの成立後は、基本的にタスクごとの活動が主であり、各リーダーのご苦労には頭が下がる思いである。心から謝意と敬意を表したい。同じタスクに興味を持っている者が一同に集まり、議論をしあつたことは非常に良い刺激

になり、そのタスクの問題や課題が明確化され、方向性を見出すきっかけになったのではないかと思  
っている。タスクによっては、物理的に集まるミーティングを数回行ったたり、メールでの議論が 100 通  
を超えたりしているタスクもある。特に多くの参加者が集まって活発に活動したわけではないタスク  
もあるが、それはそれで他のタスクとの関連性やエラー分析の重要性を認識する機会となっていれば、  
本プロジェクトの意義があったものと思う。

また、個々のタスクには直接関係しないアドバイザーの就任を以下の方々をお願いした。様々な機  
会に貴重なご意見ご指導を賜っている。ここに記して感謝の意を表したい。(敬称略)

賀沢秀人 (グーグル)

相澤彰子 (情報学研究所)

黒橋禎夫 (京都大学)

加藤恒昭 (東京大学)

奥村学 (東京工業大学)

佐藤理史 (名古屋大学)

鈴木久美 (マイクロソフト)

神門典子 (情報学研究所)

中村哲 (奈良先端大学院大学)

酒井哲也 (早稲田大学)

徳永健伸 (東京工業大学)

佐々木裕 (豊田工業大学)

影浦峯 (東京大学)

本プロジェクトは、全体調整役として、東北大学の乾健太郎、ニューヨーク大学の関根聡の 2 人で  
進めている。

## 4. 活動内容

本プロジェクトは、2014 年 3 月の言語処理学会第 20 回年次大会ワークショップ「自然言語処理の  
発展に向けた情報共有・討論」での議論を発端とし、現在のところは、2015 年 3 月の言語処理学会第  
21 回年次大会ワークショップ「エラー分析 WS」まで続いている。18 のタスクを立て、各タスクでの  
活動が中心ではあるが、全体を通じた活動の機会が 5 回あった。それらについてここに記載し、活動  
の記録とする。

### 4.1. 「自然言語処理の発展に向けた情報共有・討論」での議論

2014 年 3 月 17 日 年次大会ワークショップとして 北海道大学にて

このワークショップでは、午前中に若手の発表があり、特に飯田氏 (東工大: 当時)、松林氏  
(東北大) が、データを丁寧に作ること、データを分析することの重要性について述べた。会  
場からの活発な議論も含めて、本プロジェクトの方向性に沿った非常に有益な発表であった。  
午後の発表を予定していた筆者は、これらの発表と反響を受けて、本プロジェクトの全体調整  
役を共に務める乾氏と昼食を取りながら本プロジェクトの基本構想について話し合った。そし  
て、午後に予定されていた自身の発表までにスライドを加え、エラー分析ワークショップの構

想についての発表を行い、聴衆にその是非を問うた。それに対し、非常に前向きな意見が多く出され、この構想の意義を確認し、プロジェクトの着手を決意した。ワークショップ後、多くの人の助けを借りながら、プロジェクトの設計に入った。この北海道大学でのワークショップが Project Next NLP の生みの親であり、ワークショップを企画した岩倉氏（富士通研究所）を始めとする関係者の方々には心からの感謝を表明したい。

#### 4.2. キックオフミーティング

2014年5月19日 情報学研究所にて 約45名の参加者

本プロジェクトのキックオフを目的に、全体調整役の2名、大多数のリーダーの方、その他興味のある方の参加を得てミーティングを行った。筆者からのプロジェクトの構想の紹介、それに対する質疑応答、各タスクリーダーからのタスクの紹介があった。また、情報交換を目的に、機械翻訳におけるエラー分析の紹介をグラム氏に、情報検索におけるエラー分析の過去の事例の紹介を酒井氏に、データ共有を目的に東ロボの紹介を横野氏にさせていただいた。プロジェクトの議論として飯田氏がスライドを用意され、エラー分析の意義、アノテーションの重要性、難しさについて述べられ、プロジェクトの推進に関する提案をされた。議論の中心はプロジェクトの進め方で、「初の試みであるため、タスクの自主性に任せ、整合できるところは後から整合していこう」という立場と、「共通のデータと共通の基準を使い意味のあるエラー分析をしていこう」という立場での意見交換があった。現状では全体を俯瞰することは不可能であり、全体に共通の基準を設けることはできないという現実が認識されたものと思う。それぞれのタスクにおいて基準が違って、それらを横断的に見ることに意味があるという意見が表明された。共通のデータを使う意義についても提言があったが、実際にどういったデータを使うかはタスクの自主性に任された。既存のデータでも良く、あまり頑張りすぎずにできることをするだけでも意味があるはずである。大きくは、データの選択や内容はタスクの自主性に任せ、後ですり合わせられるところはすり合わせていこうという共通認識ができ、そのような方針がまとまった。

#### 4.3. ミッドタームミーティング

2014年9月2、3日 首都大学東京 南大沢キャンパスにて 60名を超える参加者

自然言語処理研究会との連続開催で、非常に多くの人が集まった。全てのタスクからの報告があり、全体が俯瞰できたという意味でも重要であったと思う。1日目は半日の開催で、そのうちの2時間は各タスクでミーティングを行い、それぞれ議論をしていただいた。これから活動を始めるタスクではタスクの設計、すでに活動しているタスクではそれぞれの活動の報告と意見交換などが行われた。その後、1日目に3つのタスクからの報告、2日目には残りの15のタスクの報告があった。全体の議論をした後に、懇親会としてBBQを行いながら、議論を続けた。18のタスクから各20分の発表をいただいた内容は多岐に渡り、それをこの報告にまとめるのは至難の技ではあるが、乾氏、関根がレポートをまとめ、工藤氏が参加者へのアンケートを取り集計していただいた。



## 乾氏のミッドタームミーティングのまとめ

目的を何にするか

\*エラーの全容が知りたい

- そのタスクのエラーの種類・原因について具体的な理解が深まる
- タスク間に共通のエラーの種類・原因について具体的な理解が深まる
  - タスク間の連携
    - 前段のエラーが後段にどう影響するか
    - 後段で大間違いする前段のエラーは何か
- 同じか? 裏表?
  - システムの振る舞いの分析
    - なぜその誤りを出したのか、何が足を引っ張っているのか
    - 今の手法では扱っていない現象・手がかかり
  - 改良方向の究明
    - どうすれば解けそうか? 何が手がかかりになりそうか?
    - タスクの分解(東ロボ)
    - 「なぜ人間はこれが誤りだとわかったのか」(松林、グラム)

\*分析したデータ(注釈付きデータ)を蓄積・共有

- 通り揃っていることの価値
- 誤り原因の累計を一斉公開

\*分析方略(エラー分析の方法論)をみがく

- データ、システムの準備
- ボトムアップにやってすりあわせる?
- どういう苦労があるか/ベストプラクティスの共有
- マニュアル化

\*上流と下流の違い

- 下流のエラーの一部は上流のどの工程のエラーかで分類できる
- 上流(基礎技術)のエラー分析は簡単でない

\*「なぜ人間はこれが誤りだとわかったのか」の分析が重要

## 関根のミッドタームミーティングのまとめ

- とにかく、タスクによって色々である。
- 上流のタスクは他のタスクの責任だとなすり付けることが出来ないし、言葉になかなか表現できない問題を含んでいる
- 下流のタスクは上流のタスクの融合体であることが多く、それぞれのエラーの分布や現状を見つけることだけでも非常に有意義
- でも、下流のタスクならではの深いまたは新しい課題を見つけることにも十分気を回

すことが重要

-- すでに精度がかなりサチっているタスクや標準アノテートデータがあるタスクと、新しいタスクでは、分析方法からしてかなり違う

-- エラー分析の方法論について、一般的にあるべき姿を共通に示すことは難しいし、やる必要もないだろう。ただ、列挙することは大事

-- タスク間の依存関係がある場合には、その関係を明示的に示し、それに基づく分析を行うことは非常に重要

--- 分析結果（エラーを含んだ例）を下流から上流に戻して、それに対して上流タスクがエラー分析をすることは非常に有意義

--- 上流のタスクにおけるエラーが一様に下流に影響する訳ではなく、それを調べることも重要

- エラー分析って何？

-- 1つのシステムのエラーを分析するだけでいいのか？

-- 複数のシステムの共通のエラーを分析することは理にかなっているか？

-- 凄く単純なシステムで、そのタスクに必要な技術を見極める方法論もある

-- または、全くシステムを用意せずに、そのタスクに必要な技術を調べていくという方法論もある

--- この方法はエラー分析とは言えないかもしれないが、今回のプロジェクトの目的を達成するための方法として十分

- エラー分析としてどんな種類の結果を出すか？

1) エラーの分類

2) エラーを解決する方法の分類

3) どんな現象があるのかの列挙

--上記の3つは関連があるだろうが、違ったものであろう。関連性が分かれば素晴らしい

- 分析方法論の模索？

1) 参加各人が別々にエラー分析をして、それを持ちよって議論する

2) 少人数でパイロットをして、分類のカテゴリーを決め、それに従って大勢でエラー分析を行う

-- それぞれに長所短所はありそうだが、1の方法で議論することで新しい発見はありそう

-- タスクの種類、参加者の熟練度合いによってどちらがいいかなども違うかもしれない

#### 工藤氏のアンケートのまとめ

- ミッドタームミーティングの時間配分は適切だったと考察される。若干全体討論時間が長いという意見があるが、有意差はないだろう。

- 満足度に関しては、個別ミーティング・全体会議に大きな差は見られない。全般的に高い満足度となった。

- タスク間のつながりを重視している方が多い。次回の WS では、何かつながりが生まれるような企画が求められる。

- 分析方法が確立されていないタスクが多いことが分かる。裏を返せば、これまでまともに分析ができていなかったタスクが多い？

- 正解データの再現では不十分という認識を多くの人が共有している。

- 次回 WS では、全体討論会・パネルディスカッション・ポスター・関連タスク間での中規模ミーティングが期待されている。

逆に、小規模な会議やハッカソンは人気が少ない。「タスク間のつながり」を重視されている方が多いので、

その期待にこたえられるよう、パネルディスカッション・ポスター・中規模ミーティングの開催も考慮に入れるべきであろう。

際立って言えることは、「タスク間のつながり」を重視している方が多いことです。裏を返すと、

タスクの分業化が進んでいる現状に満足していない方が多いといえます。タスク間につながりが生まれるような

メタタスクにみんなで取り組むというのもありますが、分析結果・データを参加者で共有することが第一歩だと思います。

余力があれば、別のタスクに取り組んでみることも有益かもしれません。私事で恐縮ですが、機械翻訳に取り組むことで、

現状の形態素解析や構文解析に足りないものを再認識できるようになりました。

#### 4.4. 中間レポート

2014年11月15日締め切り

ミッドタームミーティングでの熱い議論や活発な活動を維持することを目的に中間レポートの提出をそれぞれのタスクにお願いした。各タスクつき2、3名のアドバイザーに、レポートに対するコメントを依頼した。最終的には11のタスクからの中間報告があり、すべてのタスクに対して、アドバイザーからのコメントが返信された。最終的なエラー分析 WS を前に、それぞれのタスクにとって有益な方向性への指針となっていれば幸いである。また忙しい中、コメントをいただいたアドバイザーの皆様には心から感謝したい。

#### 4.5. 「エラー分析ワークショップ」

2015年3月20、21日 京都大学にて

言語処理学会のワークショップとして2日間にわたり本プロジェクトの報告会を開催する。ここでは、各タスクの1年間の活動報告と、関係タスクのディスカッション、および全体のディスカッションがメインのイベントとなる。非常に楽しみである。

## 5. 今後

本プロジェクトは、与えられた定義に基づいた与えられたタスクに対して参加者がシステムを作り、その評価を行って、システムの性能だけを比較するような評価型プロジェクトではなく、自身の作っているシステムを丹念に分析し、その経験を共有し、自然言語処理の研究者または対象のタスクの研究者が協力しあって、より良いものを作っていこうというものである。この1年の活動では、その一歩を踏み出しただけに過ぎず、このプロジェクトの形態そのものが非常に大きな可能性を秘めていると考えている。実際に、熱意を持って進めているいくつかのタスクからは、今後も是非続けていきたいという声が聞かれている。18ものタスクを平行にやる方法が良いのかどうかなど、プロジェクトの形態については慎重に検討をしながら、エラー分析のプロジェクトは今後も進めていきたいと、個人的には考えている。本プロジェクトは、エラー分析に参加している皆様が主役であり、その苦労には心から感謝を述べたい。皆様がお互いに学び合えることができ、今後の自然言語処理の発展のための一助になれば幸いである。

## 参考文献

言語処理学会第20回年次大会ワークショップ「自然言語処理の発展に向けた情報共有・討論」  
北海道大学 2014. <https://sites.google.com/site/nlp2014ws/>