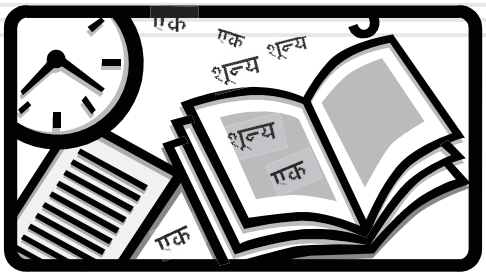

SURPRISE
LANGUAGE
EXERCISE

解説

不意打ち言語試験!?

ニューヨーク大学
関根　聡
sekine@cs.nyu.edu

米国の自然言語研究の分野で，サプライズランゲージ（Surprise Language Exercise：SLE）というユニークな実験が行われた[1]．主催したのは国防省系プロジェクトのTIDES（Translingual Information Detection, Extraction, and Summarization）であり[2]，それに参加している15程度の大学，研究所が実験に参加した（**表-1**）．

現在，アメリカ合衆国はテロの脅威に慄いている☆1．この状況下で自然言語の研究者はどんな貢献ができるであろうか．そこで練られたシナリオが以下のようなものである．現在のような世界情勢の下では，いつどこの国が戦略的に重要になるか分からない．もし，ある国で何かが起こったら，その国の情報を得ることは非常に重要である．もし，その国の言語が英語やよく知られている言語でなかったらどうするか．時間は限られている．自然言語処理の技術は使えないか．たとえば，新規の言語に対して機械翻訳システムを短時間で実現するのは可能なのか．

そのテストとして行われたのがSLEである．実際には6月2日に対象の言語が発表され，6月30日までに4つの分野，すなわち，1）機械翻訳，2）外国語のテキストから情報検索をしようという言語横断検索（Cross Lingual Information Retrieval），3）人名，組織名などの名前を抽出する固有表現抽出，そして，4）要約，のそれぞれの分野においてシステムを作成し，評価しよう，というものであった☆2．そして，2日，メールによって発表された言語はヒンディー語であった．すでに一般公開されている辞書や言語処理ツールもあるヒンディー語がサプライズなのか，という意見はあったが☆3，とにかく，DARPAからお金をもらっている大学や研究所はやるしかない．私も当日は（日本語で書かれた）ヒンディー語の教科書をニューヨークの日系書店を回って探した．

1つのグループですべてのことができるものではない．本プロジェクトは，最後にそれぞれのシステムの評価があるとはいえ協調的に進められた．テキストや辞書などのデータ集めはその道のプロであるLDC（Linguistic

検索	BBN，IBM，マサチュセッツ大学，メリーランド大学，ニューヨーク市立大学，カリフォルニア大学バークレー校
抽出	BBN，MITRE，ニューヨーク大学，シェフィールド大学，マサチュセッツ大学
要約	メリーランド大学／BBN，南カリフォルニア大学・ISI
翻訳	南カリフォルニア大学・ISI，ジョンホプキンス大学／ニューヨーク大学／メリーランド大学，IBM，カーネギーメロン大学
リソース	LDC，メリーランド大学／ジョンホプキンス大学，ニューヨーク大学，SPAWAR，カリフォルニア大学バークレー校，その他

表-1　参加団体

☆1 または，慄いているという振りをして動いている．
☆2 この試験の前に予備試験が4月に行われている．そのときの言語はフィリピンのセブ島で話されているセブラノ語．ここには米軍基地があり，政治的にも不安定らしい．
☆3 この実験のためにLDCは世界の314の言語における言語資源の存在を調べ，結果を公開している（http://www.ldc.upenn.edu/Projects/TIDES/SurveyReport.xls）．

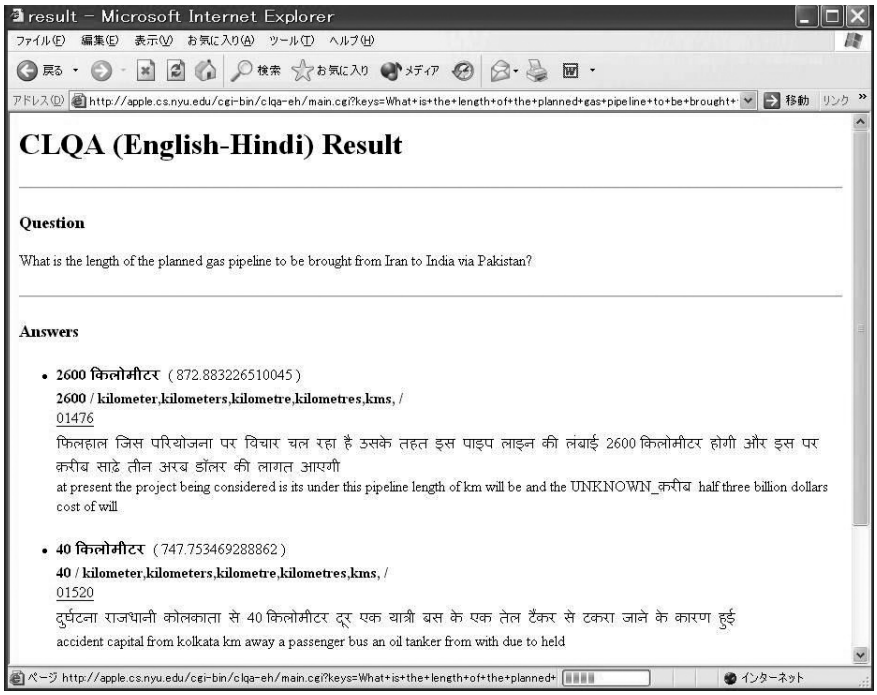

図-1　言語横断質問応答システム画面

Data Consrtium）が主に引き受け，集められた情報は参加者にWebページを通して即時公開された．共同でデータを作成しようという試みもあった．また，メーリングリストが設置され，各グループが作成したデータやツールの情報も即時流され，データやツールの多くが共有された．1カ月間で約400通のメールが流れ，各グループのリーダーは電話による会議を頻繁に行い協力体制を推進していった．と，このように和気あいあいという感じでプロジェクトは進行した．

6月のかなり最初の頃に大きなテキストデータ，良質な辞書が見つかり，開発はスムーズにいくものと思えた．が，その後，技術的な問題が立ちはだかった．それは文字のエンコーディングの問題である．ヒンディー語は，インド独特のISCII，ローマ字で書かれたITRANをはじめとしたいくつかの表記法，各新聞社固有のコード，そしてUNICODEが入り混じっており，それぞれの間のすべてをきちんと交換するようなツールが存在しなかった[☆4]．また，これは6月の最後の頃になって分かったことであるが，複数のUNICODEのコードが同じヒンディー語の文字を意味する場合があった．このように自然言語処理が研究対象とするべきではないところで大きな問題があり，時間が費やされた．

プロジェクトでは，筆者の所属するニューヨーク大学が参加した固有表現抽出と，筆者が作成した言語横断質問応答システムを中心に紹介する．固有表現抽出は他の3タスクに比べて成功であったと思われる．固有表現抽出の技術は大規模な正解データを作成し，そこから抽出のためのルールを学習するという方法が現在主流であるが，いくつかのグループが共同して正解データを作成し，そこから学習したシステムは，英語のような大文字小文字のない新しい言語においても十分に動作することが分かった[☆5]．また，筆者はこの固有表現抽出システム，CLIR，翻訳システムを利用して言語横断質問応答システムを作成した（**図-1**）．これは，英語で"Who is the prime minister of India?"と聞くと，ヒンディー語の新聞記事の中から答えを探し，それを英語に翻訳し直して"Atal Bihari Vajpayee"と答えるものである．基本的な技

☆4 もちろん，一部にはあったが，途中で不具合などが発見された．新聞社独特のものについてはコード定義さえ存在しないものもあった．

☆5 固有表現抽出は日本語，中国語，韓国語，欧米の言語などで応用に動くことは確認されているが，それが新規の言語に対しても1カ月の単位で作成可能であることが分かったのは収穫であった．

術は従来の単言語の質問応答システムに言語横断的な要素を加えたものである．作成したシステムは，デモとしても面白く，ほかに規定のタスクを大きく超えたシステムを作成したグループがいなかったこともあり，SLEのグループの中で興味を引いた．また，ここで使用している翻訳システムは南カリフォルニア大学のサーバを使っており，このようなところにも協調的なプロジェクトの雰囲気が見受けられると思う．他のタスクについては自分が深く携わったわけではないので詳しくは述べられないが，これまで知られている手法を試みただけであったという点は確実である．機械翻訳，言語横断検索については評価結果の数字が出ており，きちんとした評価までたどり着いたようであるが，要約については，ヒンディー語の要約作成の困難さ，それを翻訳する際の困難さからきちんとした評価までは至らなかったようである．詳しくは特集号が組まれる予定のTALIP（ACM, Transaction of Asian Language Information Processing）の論文誌や来年の国際学会などを注目していただきたい．

8月にプロジェクトの総括の会議があり，リーダーであったDonna Harman博士がSLEを以下のようにまとめている．綺麗にまとまっていると思うので，基本的にそのまま紹介する．

成功した点：

1) チームワークで実現させた．
2) いろいろなツールが揃った．
3) これまでの技術を新規の言語で実現することが可能だということが分かった．

失敗であった点：

1) 言語の調査が不十分であった．特にエンコーディングのような本来の課題ではないものに時間を費やしてしまった．
2) 新規のアプローチが見られなかった．

成功した点の1番に挙げられているのは，まさしくそのとおりだと思う．多くの大学や研究機関が，作成したデータ，ツールをすぐに公開し，それが他のグループで使われ，反応が返ってくるというような形式のプロジェクトはこれまで例がないのではないかと思う．また，このようにコミュニケートしていく中で仲間意識が芽生え，良いかたちでのチームワークが形成された．研究は研究者間の競争であるという点は一般的に存在すると思うが，短期間とはいえそれを超えたかたちで1つの大きなグループとしてプロジェクトが推進できたのは大きな収穫であったと思う（個人的にも米国内に知り合いが増え恩恵をこうむっている）．成功した点の2に関しては，今後，もしヒンディー語のシステムを作る際には，まず見るであろうというシステムが揃ったといえる．逆に，失敗の2点も誰もが共通に得ている感想だと思う．総括の会議でも「研究が1カ月ストップしてしまった．このようなかたちのプロジェクトは2度とすべきでない」という発言が（スポンサーを目の前にして）公然とされていた．もっともだと思う．和気あいあいと何かを作る作業は楽しい．しかし，作ったものは（DARPAの上層部を説得する材料という以外には），あまり役に立つものではない．確かに，今回の発見ともいえる「ヒンディー語の言語処理にはエンコーディングの問題があり，他の言語を対象とするときにも重要視すべき点である」ということが分かったのは収穫であったかもしれない．しかし，これだけの多くのグループの時間と労力を使わなくてもそれは分かったはずである．また，当然ながら他の言語を対象とする際には新たな問題が隠れている可能性がある．そのための効率的な方法として，たとえば，それぞれのグループがいくつかの言語を受け持ったり，問題点をあらかじめ綿密に調査するのがいいのではないか，というような意見が盛んに提案されていた．

振り返って，日本ではこのような形式のプロジェクトは可能だろうか．もちろん，米国のものをそのまま何も考えずに導入したらいいなどという馬鹿げたことを言うつもりはない．日本でも自然言語に関連した協調的プロジェクトとしてICOTやEDRがあった．しかし，これらは長期目標型であり，SLEの短期目標型とは異なる．もちろん，どちらがいいかという議論ではない．目標によって異なるものだと思う．しかし，日本ではどのような形式にしろ大学を主体とした協調体制はなかなか成立しにくいのではないだろうか．数多くの組織が協力して前に進むというのは，少なくともいくつかの利点を持っているものと思う．1つのやり方として考慮してみる価値のあるものだと思う．

参考文献

1) Surprise Language Homepage: http://www.ldc.upenn.edu/Projects/SurpriseLanguage
2) TIDES Homepage: http://www.darpa.mil/iao/TIDES.htm

（平成15年10月1日受付）