IT インダストリーレポート,
pp. 35-37, 2001. mar.

寄稿
Contribution

電子協・機械翻訳・自然言語処理

旧JEIDA 自然言語処理技術委員会 委員長
東京工業大学 大学院 情報理工学研究科 教授 田中 穂積

はじめに

思い起こすと、新しい体制に衣更えする前の日本電子工業振興協会(略称:電子協)の機械翻訳システム調査専門委員会は、わが国の自然言語処理技術発展の歩みとともにあったといえる。官学民一体となったさまざまな調査研究活動を通じて、わが国の自然言語処理技術の推進母体として大きな役割を果たすとともに、自然言語に関する技術・研究成果を世界に向けて発信する役割をも果たした。この分野の一研究者として、初期の段階からこうした活動に加わり、さまざまな研究者と議論する機会をもてたことは幸いであった。以下では、電子協と関連させつつ、機械翻訳・自然言語処理と筆者とのかかわりについて述べてみたい。

電子化文書と情報ネットワーク社会

新しい世紀を迎えた今、自然言語処理技術の背景にはこの20年間に大きな変化があったと言える。文書の電子化と情報ネットワーク社会の到来である。自然言語で記述した文書の電子化は急速に進み、今や大量の電子化文書が、家庭に、オフィスに溢れている。一方、コンピュータ技術とネットワーク技術を結合した情報ネットワークは、世界的な規模で急激に拡大しとどまるところを知らない。ネットワーク上には、大量の電子化文書が世界各地に分散して存在している。これらの文書は、家庭やオフィスに居ながらにして即座にアクセス可能になってきている。

新たな問題も発生してきた。第1は、電子化文書の量の問題である。激増する文書の中から、意図した文書を選択的に取り出す技術がまだ未熟なために、このまま放置しておくと大量の文書情報の山に埋もれてしまう恐れがでてきた。必要な時に、必要な文書にのみアクセス可能な技術の研究が緊急の課題になっている。この問題を抜本的に解決するためには、自然言語処理技術の高度化を図る必要がある。

第2は、多言語の問題である。世界にはおよそ6,000種以上の異なる言語が存在している。地球規模に拡大した情報ネットワーク上には、さまざまな言語の文書が存在している。日本語は、欧米の言語と文法が相当異なる言語である。日本人が他の言語の習得に困難を覚え、他国の人は日本語の習得に困難を覚える。機械翻訳技術が夢の技術として期待されているのはそのためであろう。ところで機械翻訳のコアとなる技術が自然言語処理技術にあることは言を待たない。

機械翻訳技術

電子協では、機械翻訳システムとその技術の重要性を見抜き、1980年代に入り、大学と主要コンピュータメーカの研究者を中心とした機械翻訳システム調査専門委員会(主査 長尾 真 京都大学教授(現京都大学総長))を立ち上げて活動を開始している。この委員会では、1980年代の初頭に、ヨーロッパ共同体参加国間の使用言語の相違を克服するための機械翻訳システム計画(EUROTRA計画)が具体的に動き出したのを契機に、調査団を世界各地に派遣して調査を行い、わが国で機械翻訳技術の研究をどう推進すべきかを提言している。その後のわが国での機械翻訳技術の立ち上がりは早く、機械翻訳の技術水準は、1980年代半ばには世界のトップレベルに達している。その意味で、電子協の果たした役割は極めて大きいものがある。

1982年から4年間、科学技術庁の支援を受け京都大学を中心に実施した機械翻訳計画は、機械翻訳システムの商用化に大きく貢献している。1980年代半ばには、いくつかの機械翻訳システムが市場に登場している。このように急速な機械翻訳技術/システムの進展とともに専門委員会では、技術動向だけでなく、市場動向、機

械翻訳システムの評価法についても検討を重ねている。

機械翻訳システムが一般に広く使われるにつれて、機械翻訳システムの不備が問題になった。特に装備されている辞書の量と質(内容)の問題が明らかになったのである。そこで通商産業省(現経済産業省)の支援を受け、1997年より9年間の予定で20万語規模の電子化辞書計画(EDR計画)が発足した。また通商産業省は、1997年より6年間、近隣諸国(中国、マレイシア、タイ、インドネシア)へ機械翻訳技術を移転する目的で、通称CICC多言語間機械翻訳研究協力計画を実施した。これは電子協の機械翻訳システム調査専門委員会の主要メンバーが中心となり、通商産業省とともに計画を策定したものである。2つの計画はともに、世界に対して大きなインパクトを与えた。

こうしたわが国での機械翻訳研究活動のうねりを受けて、電子協の機械翻訳システム調査専門委員会は、1987年に、世界中の機械翻訳システムの研究者・開発者・ユーザが一堂に会した第1回の機械翻訳に関する国際会議(機械翻訳サミット)を箱根で開催することを企画した。この会議は機械翻訳関連技術だけでなく、ユーザや政府関係者をも含む幅広い立場から、機械翻訳を議論する場を提供するものであった。その成功を受け、第2回は西ドイツのミュンヘンで開催、第3回は米国のワシントンで開催と、現在に至るまでに合計7回の会議が開催されている。次回(第8回)は2001年スペインで開催予定である。

特記すべきことは、第3回機械翻訳サミットの場で、わが国のイニシアティヴで国際機械翻訳協会(IAMT)が設立されたことである(初代会長 長尾 真氏)。そして、アジア、ヨーロッパ、北米の3地域に機械翻訳協会が設立された。アジアでは、電子協の関連協会として、アジア太平洋機械翻訳協会(AAMT)が設立された。今後この3地域が持ち回りで、2年に一度、機械翻訳サミット開催を受け持つことが決定され現在に至っている。

機械翻訳技術から自然言語処理技術一般へ

アジア太平洋機械翻訳協会の設立により、機械翻訳技術の調査研究、市場動向調査、システム評価法の検討はそこで行うことになり、電子協の機械翻訳システム調査専門委員会は機械翻訳のコアとなる自然言語処理技術一般に対象を広げ、性質・名称ともに衣更えした新しい委員会(自然言語処理技術委員会)を発足させ、活動を行うことになった。これは先に述べた、1990年代半ばから急速に発展したネットワーク社会の到来と無関係ではない。ネットワーク社会では、ネットワーク上に存在する大量の電子化文書の中から、所望の文書を取り出す文書検索技術、文書を要約する文書要約技術、内容に応じて文書を分類・整理・管理する文書分類・管理技術が重要になる。これらの技術を検討する委員会が必要になったからである。委員長が、長尾教授から筆者にバトンタッチされたことを受けて、音声と言語を統合することも、この委員会の検討課題としたいという希望を持っていたが、多くの委員の賛同が得られずに現在に至っている。

技術的展望

筆者なりに、今後重点的に研究すべき自然言語に関連した重点的な研究課題を列挙してみる。

1. 音声に関するもの

(1) 音声理解技術(自然言語処理技術と音声認識技術の統合化)、雑音に強い音声認識技術、不特定話者を対象にした連続音声認識技術
(2) イントネーションなどのパラ言語的現象の解析技術
(3) 話者認識技術
(4) 自然で滑らかな音声合成技術

2. 自然言語に関するもの

(1) 形態素・構文解析技術(文を構成する単語の認定・構文構造の抽出)
・未登録語(未定義語)の抽出、複合名詞の解析、Ill-formed文の解析(音声・話し言葉の解析)
(2) 意味解析技術
・語義曖昧性解消(機械翻訳における訳語選択の問題とも関係する)、意味を用いた構文構造の曖昧性解消、意味表現(構造)形式の確立、大規模

意味辞書・知識ベースの構築
(3) 文脈解析技術
・省略語の補強、照応関係の解析、対話(談話)構造の解析と管理
(4) 語用論的解析技術(意図解析技術)
・間接言語行為の解析、比喩的表現の解析
(5) 文章生成技術
・照応表現、省略を含む文章生成、さまざまな言い替え表現の生成、複数の文から成る物語生成
(6) 音声・自然言語解析用の大規模(構造付き)言語資源(コーパス)の構築
(7) 言語資源からの自然言語解析用知識獲得技術
(8) パラ言語的現象(身振り、手振りなど)の解析技術
(9) 音声認識と項目(2)、(3)、(4)の統合化
(10) Rule based法とcorpus based法の統合化
(11) 評価技術

解析精度については、文長や解析の深さにもよるが、新聞記事を対象にして、形態素・構文解析結果の文単位での正解率として90%を目標とすべきである。現状では正解率は50%前後である。項目(1)はすでに一定の成熟度に達している。語用論的解析技術、パラ言語的現象の解析技術、音声認識と項目(2)、(3)、(4)の統合化などは、やや困難で長期を要す研究課題である。音声認識と項目(1)との統合化や項目(10)については、すでにいくつかの試みがある。

自然言語処理技術において、項目(6)の大規模言語資源の構築が重要である。文例数が数千万規模の言語資源の構築は、さまざまな自然言語処理技術開発のための基礎データ(特にcorpus based法、自然言語解析用の知識獲得、自然言語処理技術の網羅性評価用の基礎データ)になるだけでなく、言語学の進展にも寄与する。音声認識のブレイクスルーは、大量の言語資源から得られる統計的な情報を利用したことにあった。自然言語処理技術についても同様なことが言える。ただ自然言語処理の場合には、音声とくらべてより大規模な言語資源の構築が必要になる。大規模言語資源の構築はテーマが地味であるせいか、欧米の取り組みに比べてわが国の立ち後れが目立つ。このことを危ぐし、電子協の協力を得て筆者なりに言語資源共有機構を立ち上げたが、予算の裏付けがまだ得られていない。言葉を重要視する国とそうでない国との文化的な違いが、ここに現れているように思う。「人ゲノム地図」が将来の医療産業のインフラであるとすれば、言語資源は言語産業における「人ゲノム地図」のはずである。しかし、多くの人の理解を得るに至っていないのが残念である。

3. 応用システムに関するもの

自然言語処理応用システムとして今後重点的に研究すべきものを順に列挙する。
(1) 機械翻訳システム
・制限言語、訳語選択技術、トランスファー方式、中間言語方式、多言語翻訳システム、自動通訳システム、外国語教育システム
(2) 音声理解・言語理解・対話システム
・質問応答システム、音声対話によるロボットの行動制御
(3) 文書検索システム、文書要約システム、文書分類・管理システム
(4) 文書読み上げシステム、手話生成システム、点字タイプライター
(5) 電子図書館

おわりに

自然言語処理に関する技術は、必ずしも産業界の大きな利益に直結しているとは言えない。半導体産業ほどの決定的な影響力は持っていない。にもかかわらず、将来におけるこの技術の重要性を認識し、通商産業省とともにこの技術の発展を支えてくれた電子協に対して、自然言語に関係する電子協の委員会に携わってきた一人として、深く感謝して結びとしたい。新体制での今後の一層の発展を希望してやまない。