3.4.2　知識創成コミュニケーション研究センター　言語基盤グループ

グループリーダー　　鳥澤健太郎　ほか 18 名

用例ベース、辞書等の言語資源構築及び知的自然言語処理システムの研究開発

【概　要】

言語基盤グループは、ナチュラルコミュニケーション技術の開発の一環として、言語資源プロジェクト、言語グリッドプロジェクトの 2 プロジェクト体制で、音声・言語処理の基盤となる、大規模な言語資源の構築・公開、及びその作成・活用に資する言語処理技術、それらを統合しサービスとして実現する言語グリッドの開発を行っている。

【平成 22 年度の成果】

平成 22 年度には、音声言語技術の普及を目指して設立された高度言語情報融合フォーラム（ALAGIN、http://www.alagin.jp/）において配信するため、大規模言語資源、言語解析ツール、言語資源を自動構築するためのツール、言語資源を活用するサービスの開発を行い、また、それらを活用する知的言語処理システムの開発を行った。

【言語資源プロジェクトにおける成果】

(1)　音声質問応答システム「一休」の開発

スマートフォンに入力された音声での質問にほぼリアルタイムで回答を列挙するシステム「一休」を開発した。「一休」は Web 6 億ページから質問への回答を抽出する（図 1）。この回答抽出は人の常識、常識的行動に関する辞書中の知識を用いて、柔軟に質問に回答をするものであり、対話システムのコンポーネントとなることを念頭に開発されたものである。また、これはこれまでに蓄積された NICT の言語処理技術、特に後述する概念辞書の自動構築技術、音声認識技術、言語資源の優位性を検証する目的もあって、短時間、ローコストで開発する事を主眼に開発した。実際にそれらの蓄積をフルに活用して構築し、4 名の研究者で開発開始から 3ヶ月で日本を代表する ICT 関連の展示会「CEATEC JAPAN 2010」でデモを実施するレベルに達した。

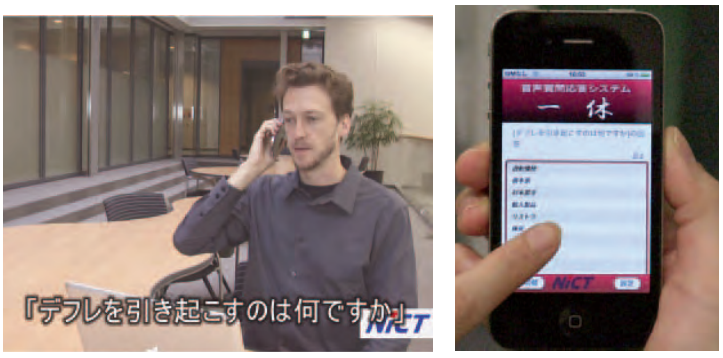

図 1　音声質問応答システム「一休」
（左： スマートフォンへの音声入力、右： 回答の提示）
デモビデオは http://www2.nict.go.jp/x/x161/ にて視聴可能

このように非常に短期間に構築されたシステムであるが、処理対象が 6 億ページの Web 文書と大量であることもあり、簡単な質問によって非常に意外でありながら、有用な回答を得られる。例えば、「デフレを引き起こすのは何ですか」といった質問の回答としては、意外なことに日本を代表する自動車メーカーの名前が提示された。これはあるブログ中に「＜その企業＞が、巨額の利益を内部留保にまわしたため、総需要が縮小し、デフレを悪化させた」という記述があったのをシステムが発見したからである。この回答を発見するプロセスを先にのべた「CEATEC JAPAN 2010」でデモした後、ある著名な経済雑誌でほぼ同主旨の記事が掲載された。つまり、発端はブログ記事に過ぎなかったわけであるが、経済雑誌で取り上げられるほどの信憑性、インパクトがある回答を先取りで発見したということになる。なお、質問は「デフレを引き起こすもの」を問うていたのに対して、ブログ中の記事では「＜その企業＞がデフレを悪化させた」と記述されていた。一休を支える技術の代表的なものが「言い換え」の自動認識技術であり、この技術により「引き起こす」と「悪化させる」という字面上全く異なる表現が非常に類似した意味で使われているということが自動的に認識された。こうした技術およびそうした技術によって作成できる辞書は、人の常識、常識的行動を捉える上で非常に重要であり、今後さらなる展開を計画している。また、一休には、Web 上に書かれた知識に関する推論技術が導入されており、そもそも入力となる Web 文書に（少なくとも直接的に）書かれていないが、妥当である可能性が高い知識を質問の回答として提供することが一部可能になっている。今後こうした推論技術をさらに拡張させることによって、Web を単に多様な情報が記載された情報源・デー

タから、いわば「考える主体」へと進化させることが可能となろう。

また、一休は 6 億ページという大量の Web 文書から回答を発見するにも関わらず、サーバー1 台でほぼリアルタイムでサービスを実施する事ができ、若干大きめのハードディスクを搭載した PC を自宅に所有しているユーザであれば、誰でも音声による質問によって前述したような意外でありながら有用な情報の発見を行うサービスを自宅で立ち上げることすら可能である。

(2)　言語資源、言語解析ツールの構築と配信

当プロジェクトで継続的に開発している概念辞書、つまり、単語と単語の間の意味的関係を記述した巨大なネットワークであり、億単位の Web 文書から自動的に抽出・構築される辞書を拡張し、そのカバーする語彙数を平成 21 年度の 220 万語から 250 万語まで増大させた。この概念辞書に関連するデータやサービス、すなわち言語資源、言語資源構築サービスの ALAGIN での利用許諾件数は平成 21 年度の 140 件から 458 件へと増大した。(これらの言語資源の詳細については http://nlpwww.nict.go.jp/corpus/resources.html を参照されたい。) また、概念辞書の開発、特に概念辞書を Web から自動的に構築する一連の技術に関しては学術的にも非常に高く評価され、グループリーダーの鳥澤健太郎が日本学術振興会賞を受賞した。

また、概念辞書の他に平成 20 年度、平成 21 年度に引き続き、中国語の構文解析で世界最高性能を達成し、その成果である構文解析器は ALAGIN フォーラムで公開されている。また、ALAGIN フォーラムとは別個に言語翻訳グループと共同で Wikipedia 日英京都関連文書対訳コーパスなどの creative commons での公開を行った。また、平成 22 年度以前から公開を開始している日本語 WordNet のアップデートも行い、これのダウンロード件数は 8,000 件を越えている。

【言語グリッドプロジェクトにおける成果】

当プロジェクトでは、言語の壁の克服に向け、インターネット上の言語資源を連携させ多言語サービスとして提供する「言語グリッド」、およびそれを利用した多言語コラボレーションツールの研究開発を行っている。

平成 22 年度の研究成果は以下の通りである。

(1)　言語グリッドの連邦制運営の実現

個別に運営された言語グリッド間の接続を実現し、互いの言語グリッドのユーザが、多言語サービスを共有し相互に連携して利用することを可能にする言語グリッド連携基盤を開発した。これにより、これまで実施してきた京都大学大学院情報学研究科社会情報学専攻による国内の運営だけでは、言語資源提供者にアクセスすることに限界のあった海外の地域でも言語グリッドの運営組織を立ちあげることで、多様な言語のサービスの拡充が可能になる。実際に、タイの国立電子・コンピュータ技術研究所（NECTEC）が言語グリッドを運営し、バンコクの言語グリッドと京都の言語グリッドを相互に接続して言語グリッドの連邦制運営を実現することで、新たに東南アジアの言語を中心とした 13 カ国語 21 言語サービスが京都の言語グリッドユーザにも利用可能になり、東南アジアの言語サービスを拡充している（図 2）。

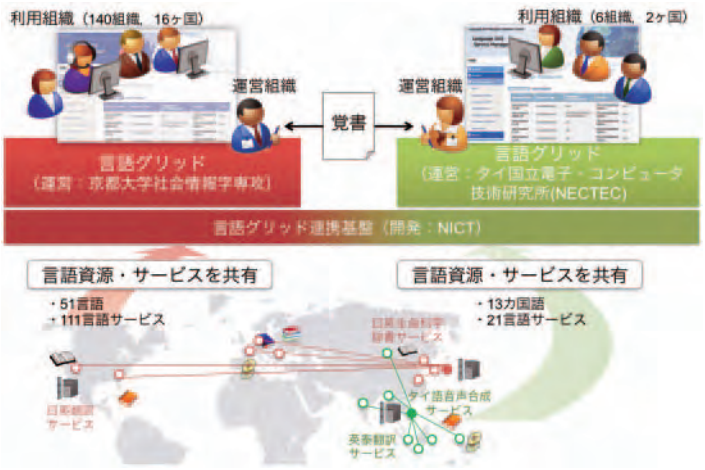

図 2　言語グリッドの連邦制運営

(2)　多言語コラボレーションツールのオープンソース化とクラウドサービスの提供

当プロジェクトで平成 21 年度に開発した言語グリッド Toolbox を多言語コラボレーションツールフレームワークとして整備しオープンソース化を行った。さらに、オープンソースプロジェクトを立ちあげ、15 組織 30 名からなるメンバでオープン型のソフトウェア開発を促進している。実際に、15 種の多言語コラボレーション支援モジュールが言語グリッド Toolbox のプラグインとして構築され、その組み合わせにより利用者コミュニティのニーズに合わせたツールが実現されている。さらに、開発した多言語コラボレーション支援モジュールを追加した言語グリッド Toolbox のクラウドサービスも運用しており、東京外国語大学や京都大学など 30 組織に利用され、留学生支援や国際交流活動などに活用されている。