3.5.2　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 31 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 21 年度の成果

【言語資源プロジェクトにおける成果】

平成 21 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、それらの実用化に向けての活動を行った。

①規模言語資源、言語解析ツールの構築と配信

平成 21 年度は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を拡張し、そのカバーする語彙数を平成 20 年度の 180 万語から 220 万語まで増大させ、概念辞書の一部である 6 種の日本語に関するデータを言語資源として ALAGIN にて配信を開始した。さらに、言語翻訳グループと共同で対訳コーパスの配信に向けて準備を進めている。これらの詳細は http://nlpwww.nict.go.jp/corpus/resources.html で知ることができる。また、Wikipedia から語とその上位概念との関係を 100 万個オーダーで自動的に抽出するツールを Web 上で一般向けに公開した（http://nlpwww.nict.go.jp/hyponymy/index.html）。こうした成果は、例えばニフティ株式会社において「@nifty 温泉」で活用されている。さらに言語解析ツールの開発に関しては、平成 20 年度に引き続き、タイ語、中国語に関して形態素解析、構文解析で世界最高性能を達成するなど、国際学会における性能比較のコンテストにおいて多数種目で、優勝もしくは、入賞した。これらのツールも公開予定である。

②言語資源構築ツールの研究開発

当グループの言語資源の目玉である概念辞書は、前述したように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るが、それらの拡張の加速、あるいはニーズに合わせたチューニングを可能にするため、研究者以外のユーザがローコストで拡張するための 2 種の Web サービスを開発した。1 つ目は、単語の意味クラス、例えば「日本酒の一覧」といったものを、ユーザからの入力をもとに自動的に構築するサービス（図 1）である。このサービスでは、Web 上の頻度上位 1,000 万語を対象に、ユーザが入力する語を出発点として、それに意味的に類似する語を単語クラスに含まれるべき単語の候補として提示し、さらにユーザからのフィードバックを繰り返すことで、ローコストで単語クラスを作成することができる。例えば、1 名の作業者が 1 日程度の作業で 6,379 語からなる「食材」の単語クラスを語の全数チェック込みで作成することができた。通常それだけ大量の語を「思いつく」ことは非常に難しいが、このサービスのポイントは、思いもつかない単語を候補として提示することにある。一旦単語が提示されれば、それが目標とするクラスに含まれるか否かの判断は遥かに容易である。

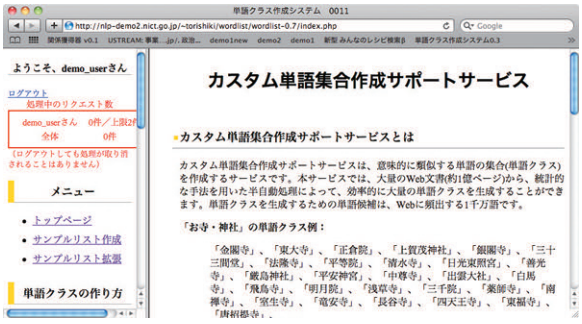

図 1　単語の意味クラスを作成するサービス

開発したもう 1 つのサービスは、単語間の意味的関係を大量の Web 文書からローコストで自動的に抽出する「意味的関係抽出サービス」である。例えば、因果関係を持つ単語の対、例えば「ウイルス」と「風邪」といったものを取得する場合には、「A が B の原因である」というような変数 A、B を含むパターンを複数個入力する。開発したサービスは A、B にマッチする単語の対を万のオーダーで Web 文書 6 億ページ

から 1 時間程度の計算で自動的に抽出する。この際、入力として与えられたパターンと同義なパターンを用いての抽出も行う。例えば、「A が B の原因である」という入力が与えられた場合、「A が B の引き金である」といった同義なパターンを自動的に発見し、それらも用いて単語の抽出を行う。これにより「場所とその名物、名所」「食材とその健康効果」など、これまでの既存研究では考慮されたことの無かった意味的関係を容易に概念辞書に加えることが可能となり、通常の検索エンジンでは見つけることのできない意外でありながら有用な情報を容易に発見することが可能となった。

図 2　みんなのレシピ検索

これら 2 種の言語資源構築ツールは平成 22 年度に ALAGIN にて公開すべく作業中であり、特に意味的関係抽出サービスは公開されるものとしては世界初となる。また、すでにこれらのサービスは、ニフティ株式会社からの受託研究による「@nifty みんなのレシピ検索」（図 2）の開発において活用されている。

【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 21 年度の研究成果は以下の通りである。

①言語サービスの開発

言語グリッド上では、複数の言語資源（辞書や翻訳ソフトウェアなど）を組み合わせることで、新しい複合的な言語サービスを構築することが可能である。これらのサービスはサービス利用者に付加価値を提供することができるが、一方で全てのサービスが実行できないと結果を得ることができないという問題も引き起こしている。このような問題に対処するために、複合サービスの実行時制御を行うサービススーパビジョンを提案している。この技術により、サービスを制御するメタなサービスを記述することができ、実行時の代替サービスへの動的な切り替えやサービスの再試行といった適応技術が可能になっている。また、新たな試みとして、プログラムやデータといった言語資源だけでなく、人もサービスとして扱うことで、人と言語資源の連携も実現している。具体的には、マニュアルのローカリゼーション翻訳作業において、翻訳ソフトとネイティブの英語チェックおよび修正をサービス化し、連携させることで、翻訳家だけでローカリゼーション翻訳作業を行った場合と翻訳品質を変えることなく、翻訳のコストを削減できることを実験により示した。

②多言語コラボレーションツールの開発

多言語コンテンツを管理するコンテンツマネージメントシステムをベースに、多言語コラボレーション支援ツール「言語グリッドツールボックス」を開発し、オープンソースソフトウェアとして公開した。言語グリッドツールボックスは、言語グリッド上の言語サービスを利用するための言語サービス設定機能や、メンバ管理、ファイル共有管理といった多言語コミュニティを支援するための機能を提供するフレームワークである。オープンソースソフトウェアとして公開されることで、このフレームワークを用いた多様な多言語コラボレーションモジュールの開発が可能になっている。具体的には、これまで京都大学や京都市により、多言語テキスト翻訳モジュールや、多言語掲示板モジュール、多言語辞書作成モジュール、多言語 Web 翻訳モジュール、多言語 Q&A サイトモジュールが開発されている（図 3）。なお、このシステムのプレスリリースを受けて、新聞、テレビ、Web ニュースなど様々なメディアに取り上げられ、15 件の報道が行われた。

図 3　言語グリッドツールボックスの多言語掲示板モジュール