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19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、言語障壁を起因とするデジタルデバイドの解消を目指し、今後の音声・言語処理の基盤となる大規模な言語資源の作成・公開及びその作成・活用に資する言語処理技術や応用システム、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20年度の成果

⑴　高度言語情報融合フォーラムの設立

音声･言語資源分野の研究開発を推進する「MASTARプロジェクト」の開始に伴い、産学官の連携により研究開発と成果の普及展開を進めるためにMASTARプロジェクト全体が企画に携わる「高度言語情報融合フォーラム（ALAGIN-Advanced Language Information Forum）」（http://www.alagin.jp/）が設立された。会長は辻井潤一東大教授、副会長は喜連川優東大教授と松島裕一NICT理事であり、60社を超える企業及び60名を超える大学関係者がメンバーである。フォーラムの目的は、産官学を巻き込んで、一層の技術共有、リソースの共有を図り、研究の効率化を図るとともに、出口としての将来のアプリケーション像を明確にするための議論を行うことである。

こうした活動の一環として、言語基盤グループでは、音声･言語資源配信サイト（http://nlpwww.nict.go.jp/forum/）を立ち上げ、音声・言語資源とツールをフォーラム会員で共有するアクティビティを開始した。具体的に平成20年度に作成し、平成21年度に配信を予定しているものは以下のとおりである。

①　用例ベース（対訳コーパス）（約100万文）
②　対訳辞書（約50万文）
③　音声対訳コーパス
④　文脈類似語データベース（約100万語）
⑤　動詞含意関係データベース（37,000対）
⑥　上位下位関係抽出ツールキット
⑦　上位語オントロジー
⑧　負担・トラブル表現リスト（約2万文）
⑨　中国語形態素解析器・構文解析器

このうち、音声対話コーパスは音声コミュニケーショングループが構築している。その他の言語資源、ツールについては、下記⑵を参照されたい。

⑵　言語資源の構築及び知的自然言語処理システムの研究開発

①　用例ベースの構築

用例ベースは機械翻訳性能の向上のために必須の言語資源である。用例ベースとしては、言語翻訳グループと共同で、合計150万文対を構築した。内訳は、京都観光情報を対象に人手による翻訳50万文、既存用例ベースに対して言い換えを適用して新たに自動生成した50万文、ソフトウエアLinuxやインターネット標準文書RFCに関わる複数の翻訳者コミュニティ作成の散在しているWebデータから自動抽出した50万文からなる。これにより平成19年度までの成果と合わせて日本語に関しては前例を見ない合計584万文対の用例ベースを構築したことになる。これらは、著作権等の権利関係の問題が解消されたものから順次、高度言語情報融合フォーラムにおいて公開する予定である。

②　辞書の構築

NICTでクローリングした億単位のWeb文書から自動獲得したものをベースに言語辞書の構築を進めた。対訳辞書として50万語規模のものを機械学習やパターンマッチングによって新規に構築した他、日本語に関する概念辞書のカバレージを平成20年度頭の130万語から180万語へ（上位下位関係）、50万語から100万語へ（上記⑴の文脈類似語データベース）と世界最大規模へ拡張した。文脈類似語データベースは、約100万語の名詞に対して、Web文書上での文脈が類似している名詞を類似度とともに順に列挙したもので、

高精度な類義語データベースとして利用でき、例えば、地魚の一覧を含むクラスなどが入手できる。上位下位関係については、上位下位関係抽出ツールを開発、公開（http://nlpwww.nict.go.jp/hyponymy/index.html）し、上位語オントロジーも整備した。上位語オントロジーは、上位下位関係抽出ツールの出力を補完するものである。ツール自体はWikipediaから100万語以上をカバーする上位下位関係を抽出するが、この上位語オントロジーを併用することでより高い精度の上位下位関係が入手できる。

また、新たに因果関係、含意関係等の新規な単語間の意味的関係をWeb文書から自動獲得し、動詞含意関係データベース37,000対（約7,000対の正しい含意関係の他、機械学習の負例用に含意関係にない動詞対も含む）、上位下位関係100万対、負担・トラブル表現リスト約２万を人手で検証した。動詞含意関係データベースは、含意関係が成立している動詞のペアを辞書順に列挙したもので、動詞１が動詞２を含意するとは、動詞１が成立するなら、動詞２も成立しているということを意味する。例えば、「試乗する」は「運転する」を、「挑戦する」は「チャレンジする」を、「チンする」は「加熱する」を含意する。負担・トラブル表現リストは、災害や病、障壁や規制など、人間の活動に負荷を与えたり、マイナス効果をもたらしたりする事物に関する表現を集めたものである。このリストは、我々が開発した検索支援システム「鳥式改」（鳥澤他、情報処理学会学会誌「情報爆発特集号」、2008年８月）でも利用されており、これによって意外なトラブルを網羅的にネット上で検索することが可能となる。

さらに、英語版概念辞書の開発にも着手し、330万語をカバーする上位下位関係データベースを構築したほか、日本語WordNetの開発を昨年度に引き続いて行い、規模は約８万語となった。日本語WordNetについては、一般公開（http://nlpwww.nict.go.jp/wn-ja/）後、多数のダウンロードが行われ、活用ツールが国内外で開発されている。

図１　一般公開中の日本語WordNet（http://nlpwww.nict.go.jp/wn-ja/）の例（左）とウェブインターフェース（右）

③　知的自然言語処理システムの開発

知的自然言語処理技術の基盤となる形態素解析に関しては、日中タイの各言語に関してstate-of-the-artの精度を達成し、構文解析に関しては中国語で世界最高の精度を達成した。それぞれ機械学習に基づく手法を採用しており、形態素解析器、構文解析器は、モデルとともに、高度言語情報融合フォーラムで公開する予定である。タイ語に関しては、平成21年２月６～８日に開催された単語分割に関するコンテスト（Benchmark for Enhancing the Standard of Thai language processing (BEST) 2009）にKasetsart UniversityのChuleerat Jaruskulchai准教授と共同で参加し、優勝を果たした。このコンテストには大学、企業から20チームが参加し、決勝に残ったのは６チームであった。強豪がひしめく中での価値ある優勝である。（図２）

知的自然言語処理技術としては、概念辞書を用いてWeb上の情報をアナロジーによって検索するシステム「鳥式改」（図３）の開発を行い、リスク管理、イノベーション支援において有効であるとの示唆を得た。具体的には、社会的にインパクトを持ち得る意外なトラブルやネットのいわゆる暗部での意外な議論、情報を多数発見することに成功している。こうした成果は昨今のWebの急激な普及、いわゆる情報爆発に対処する上で非常に重要な技術である。

図2　タイ語の単語分割に関するコンテスト（BEST 2009）で優勝したときの様子

図3　概念辞書を用いてWeb上の情報をアナロジーによって検索するシステム「鳥式改」の実行例

これらの中で、特に知的自然言語処理として開発した鳥式改は、これまでに無い形態のアプリケーションである。これまでの自然言語処理研究では、人の持つ知識を計算機上に再現するということに主眼がおかれており、実際に計算機上に再現することができた知識、情報の量は一個人の持つ知識の範囲内である場合がほとんどであった。しかしながら、概念辞書中の構造化された情報はすでに一面において一個人の持つ知識を遥かに凌駕しており、それによって一個人からすれば思いもよらないような情報の発見を可能にしている。こうした事態は、これまでの言語処理研究ではあり得なかった事態であり、今後の重要な研究の方向性を示唆するものと思われる。つまり、Web上の情報を適切に構造化し、辞書のように体系化することで、今までに無かった価値を発見、創出することが可能になるということである。今後は、こうした方向をさらに追求するとともに、他グループで開発を行っている機械翻訳や対話システムでの言語辞書の有効活用に関しても研究を行う。

(3)　言語グリッドの研究開発

言語グリッドプロジェクトでは、言語の壁の克服に向けて、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」の開発及びそれを利用した異文化コラボレーションツールの研究開発を行っている。平成20年度の研究成果は以下のとおりである。

①　言語サービスの開発

複数の言語資源を組み合わせた言語サービスとして、ドメインに特化した翻訳品質のカスタマイズが可能な機械翻訳･辞書連携サービス、HTMLコンテンツの翻訳が可能なWeb翻訳、人手で翻訳した用例対訳テンプレートを用いた翻訳が可能な用例対訳・辞書連携サービスなど新たに13種類の言語サービスを構築し、運用中の言語グリッド上で公開した。また、言語サービス実行時に利用する言語資源の切り替えを可能にする動的バインディング機能を開発した。これによりユーザによる言語資源の組替えが可能になり、サービスの多様性を実現している。一方、言語資源の連携に関する基礎研究として、複数の機械翻訳を結

合する際に生じる訳語のドリフトの問題に取り組んだ。訳語のドリフトとは、複数の機械翻訳を連携させた際の中間の結果に多義語が含まれることで、それ以降の翻訳結果が原文と異なる意味になる現象である（図４）。この訳語ドリフトを防止するために、訳語選択情報を文脈とした機械翻訳間に文脈を伝搬させるサービス連携技術の研究を行った。具体的には、既存の複数の対訳辞書から、多言語の同義語集合を獲得し、各翻訳サービスは文脈を基に多言語同義語集合を参照し、一貫した訳語選択を行う。この技術は、現在、国内特許及び国際特許（アメリカ、中国、欧州）に出願中である。

②　多言語コミュニケーションの分析

機械翻訳の多言語コミュニケーションへの適用に向けて、これまで行われてきた二者間でのコミュニケーションの分析を拡張して、三者間でのコミュニケーションの分析を行った。その結果、三言語による三者間でのコミュニケーションでは、機械翻訳を用いた二者間の三つの通信路が独立に確立されているため、参加者は他の二者間でどのような情報が共有されているのかが分からず、参加者は他者に合わせた発言を行うことが困難な傾向にあることが分かった。また、単言語によるコミュニケーションの場合、対話が進むにつれて短くなる傾向にある参照表現が、機械翻訳を用いた場合、三者間での情報共有が促進されず長くなる傾向にあることも分かった。これらの結果は、コンピュータ・ヒューマンインタラクションの分野のトップカンファレンスであるCHI09で発表予定である。

③　多文化共生活動の支援

2007年12月より非営利利用を対象に運営が開始された言語グリッドは、大学、研究機関、NPOなど12か国87組織が参加して国内を中心に利用が始まっている。教育分野では、NPOパンゲアが韓国UNESCOと連携して子供たちの国際交流支援に言語グリッドを利用しており、関西大学や立命館大学は川崎市立富士見中学校や宇治市立南宇治中学校での多言語対話支援に利用している。また、医療分野では、和歌山大学が言語グリッド上の医療用例対訳サービスを利用して医療受付対話支援システムM３を開発し、京都市立病院や京都大学付属病院に設置している。これらの言語グリッドを利用した研究成果は、「言語グリッド」をテーマにした電子情報通信学会の研究会で報告（９大学とNICTから19件の研究発表）され、言語グリッドを利用した活動の広がりを示している。

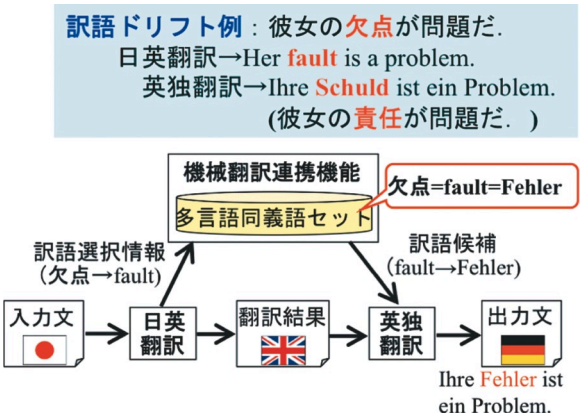

図４　訳語ドリフト例と機械翻訳連携技術

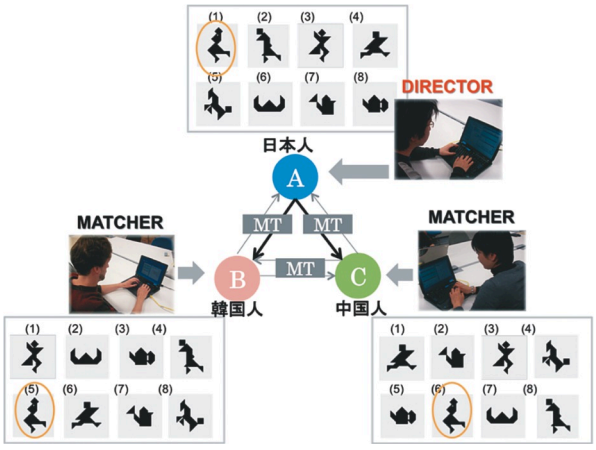

図５　機械翻訳を用いた三者間コミュニケーション実験