日中・中日言語処理技術の開発研究

実施予定期間：平成 18 年度～平成 22 年度

研 究 代 表 者: 井佐原　均（(独) 情報通信研究機構知識創成コミュニケーション研究センター)

Ⅰ．概要

アジア諸国の言語のうち本研究では、中国語に焦点をあて、大量の日中対訳コーパスを蓄積し、言語の構造を用いて適切な用例を選択・利用する技術の開発を行う。また対訳コーパスから半自動的に科学技術分野の日中・中日翻訳・情報検索用辞書を作成する手法を確立する。さらに他の言語にも適用が容易な用例翻訳手法を確立し、翻訳システムプロトタイプを開発すると共に、実証実験・評価を行う。

1. 研究の目的

科学技術面で飛躍的な発展を遂げつつある中国をはじめ、アジア諸国内において流通している科学技術情報の日本国内での活用を容易にし、かつ我が国が最先端を担っている科学技術分野の文献情報の各国への流通を促進することにより、アジア諸国と日本の科学技術の発展に資する。

2. 国内外の研究状況、提案にいたる準備・調査等について

a. 国内外の研究状況

(1) 言語解析、翻訳システムの構築

中国においては、中国科学院傘下の研究所（計算技術研究所、自動化研究所、ソフトウェア研究所など）において、言語処理及び機械翻訳の研究が行なわれており、清華大学、北京大学、南京大学等においても研究が行なわれている。

わが国においても、公的研究機関である独立行政法人情報通信研究機構（以下 NICT）やいくつかの大学で基盤的な研究が行なわれ、企業でのシステム開発が行なわれている。しかしながら、基本的な言語解析技術においては、日本語に関する技術は実用レベルに達しているが、中国語の場合はまだ精度が低い。このため、中国語を入力とする翻訳システムは精度が低く、実用的なシステムの実現には達していない。

中国国内においては、華建などの中国系企業において、実用システムの開発・販売が行なわれおり、さらに北京オリンピックや上海万博に向けて、機械翻訳をも含む研究開発プロジェクトが立ち上がっている。ただ、これらは対話文などの短い文を対象としており、比較的長文が多い科学技術に関する情報流通の翻訳研究は少ない。わが国においても、長文を対象とする実用的高性能機械翻訳を目指すプロジェクトは存在しない。

(2) 辞書構築

日本国内においては、NICT が英語を介した日中対訳辞書の自動構築手法を提案しているが、同義語、異表記語を含めた科学技術分野の数百万語レベルでの半自動構築の手法については実現していない。

独立行政法人科学技術振興機構（以下 JST）では日英・英日機械翻訳システム用の開発を 1980 年代より継続して実施中。また情報検索用日英対訳大規模辞書（同義語・異表記語辞書）の構築を平成 16 年度以降実施している。

b. 提案にいたる準備・調査等

(1) 提案にいたる準備・調査

・NICT においては、5 年間に渡り日中自然言語処理共同研究促進会議を開催し、日中言語処理に関わる共通の問題点の検討を行い、かつ研究者との交流を深めている。当会議の参加機関は日中双方とも毎回 10 機関を超え、日本及び中国における中心的な自然言語処理研究グループをほぼ網羅している。

・JST では既に科学技術分野における日英機械翻訳システムの開発および辞書の拡張を実施している他、平成 17 年 9 月に日中科学技術情報意見交換会を実施し、中国国内の枢要情報機関約 20 機関の担当者と科学技術情報流通の方策について討議し、日中・中日機械翻訳システムのニーズについても把握済み。また NICT と JST は共同で、平成 17 年 10 月に中国語言語資源に関わる調査団を派遣し、中国国内における科学技術関連の言語資源について、その種類、入手経路を調査済みである。

(2) 提案の基礎となる研究の内容及び成果の概要

・NICT は日英の文章間の対応付けを高精度に行なう技術を開発した。本技術は日中の文章間の対応付けに容易に拡張でき、対訳コーパスの作成及び用例翻訳の基盤技術として利用できる。

・NICT はコーパスからの学習に基づく自然言語処理技術に秀でており、トップクラスの精度を誇る日本語の形態素解析技術・構文解析技術・固有表現抽出・情報検索技術などを開発した。これはコーパスからの完全自動学習によるものであり、中国語などの他言語へも展開可能である。これらの技術を今回開発する中国語処理技術の基盤とする。

・NICT においては、詳細な言語情報を付与した 4 万文規模の日中英の小規模対訳コーパスがほぼ完成している。この情報付与を効率化するための編集ツール群を開発した。また、NICT が著作権を有する日英電子化辞書を拡張し、40 万語レベルの日中英の電子化辞書の開発を行なっている。これらの成果は本研究開発に活用する。

・京都大学においては、既に文の構造情報を利用する用例ベース翻訳エンジンを開発しており、日英翻訳のプロトタイプシステムも構築している。旅行対話ドメインの日英翻訳につ

いては国際的評価型ワークショップにおいて上位の成績をおさめている。

・東京大学においては、主として英語を基本として、文法記述枠組みHPSGと高効率な処理系の研究、機械学習の言語構造処理への適用、生命科学テキストのText Mining、テキストからの知識獲得、テキスト・クラスタリングなどの研究が進められており、これらの成果は日中大規模科学技術用語辞書の半自動構築に活用が可能である。また自然言語処理、情報検索、WWWについての研究を通じて多言語対応の用語抽出システム「言選Web」などの成果を有しており、アジア多言語での情報検索の基盤技術を活用可能である。

・静岡大学においては、平成17年度より英語を基本として文脈に応じて単語の意味を訳し分ける語義ネットワークの研究を行い、訳し分けの機能を有する辞書を構築中であり、高精度な訳質を提供する翻訳用科学技術用語辞書の構築にその成果を活用する予定である。

・JSTにおいては、約64万語におよぶ日英機械翻訳用対訳辞書、データソース140万語におよぶ日英対訳の情報検索用大規模辞書（同義語・異表記語辞書）を既に作成、保有しており、またその作成ノウハウも保持している。また総数800万文献分の日英対訳のパラレルコーパスも保有しており、これらの成果は本研究開発に利用する。

3. 研究内容

a. 日中、中日の用例ベースの翻訳システムの研究開発（サブテーマ1）

（1）解析システムに関する研究開発

これまで英語や日本語の解析で有効性が確認されてきたコーパスに基づく解析手法を中国語に適用し、中国語の形態素解析・構文解析の高度化を図る。

（2）翻訳エンジンに関する研究開発

用例ベース翻訳において用例を柔軟に利用できるようにするとともに、対訳コーパスにおいて語・句を高精度で対応付ける手法を確立する。

b. 日中・中日言語資源の構築と構築技術に関する研究開発（サブテーマ2）

（1）日中・中日翻訳用大規模辞書の構築

・日英(及び英中)の専門用語・一般用語対訳辞書をもとに初期版の日中・中日辞書を作成し、さらに当該辞書を活用し、基本用語と科学技術用語約200万語（同義語、異表記語を含む）の半自動収集を行い、対訳関係を収集することで機械翻訳のための大規模辞書を作成する。

・さらに、上記辞書を用いて高精度な文解析、翻訳処理に必要な意味関係の抽出、および、多義語・多訳語に対処するために必要な個別の語ごとの統計的なプロファイルを持った辞書を構築する。

・分野依存的に関連専門用語の使用パターンを統計的に分析することで、複数の用語間に現れるより豊かな意味関係を抽出し辞書に付加する手法を確立する。

・文法規則解析用に、文献データベース中に存在する中国文献のタイトル、抄録および教科書などの各種の科学技術関連文書より、最低100万文規模の大規模な日中英文献コーパスを作成し、このコーパスから科学技術文献対応の語義関連ネットワークを生成し、得られた語義関連ネットワークと訳語選択プログラムを用いて、評価例文に対する訳語選択を実行し訳語選択の精度を評価する。

・日中・中日の大規模なコーパスを収集とともに、バランスのとれた収集方法や、対訳文対を効率良く増やす手法を確立する。

c. 翻訳システムプロトタイプシステムの開発および実証実験（サブテーマ3）

上記で得られた研究成果をもとに、日中・中日機械翻訳プロトタイプシステムを作成し、実用レベルに近い機械翻訳が実現可能であることを示す。

4. 政策目標の達成への寄与、経済社会への波及効果について

本研究の進展によって、言語障壁により中国国内のみで流通している有益な科学技術情報を、本研究の成果である翻訳システムを利用し、我が国の研究者・技術者、事業者が容易に活用することにより、共同研究事業の設立など大きなビジネスチャンスを生み出すことも可能となるため、国力の進展が期待できる。また日本が最先端を担っている科学技術文献が中国国内で流通することにより、中国の科学技術発展への寄与が可能となる。

さらに、本研究の成果はコーパスを変更することにより、翻訳エンジン部分の大規模な改造を行うことなく、広くアジアの多言語への応用が可能な共通基盤となり得るため、将来的には中国のみならずアジア諸国の言語体系への適用が可能であり、アジア諸国の科学技術発展への寄与も大きいものと推測される。また翻訳システムの開発と併せて、中国語をはじめアジア多言語の科学技術情報を母国語で検索可能とするための大規模科学技術用語辞書構築の基盤技術が確立されるため、アジア各国内で流通している情報の検索を容易にすることが可能である。さらに、本システムを広く民間に開示することで当該システムの更なる進展を図ることにより、日英・英日機械翻訳システムと同等の翻訳市場を開拓することが可能になると期待される。さらに、民間企業が、本システムを広く活用することにより、アジア諸国において活躍しやすくなり、その結果、民間企業間の交流が進展し、日中間およびアジア諸国間の科学技術・経済協力にも繋がることが期待できる。そのため、経済社会への波及効果は大きい。

5. 研究終了後の実用化等に向けた自立的な取組について

a. NICTでは下記の自立的取り組みを行う。

・翻訳エンジン部分の洗練を継続して行う。

・タイ語等、他のアジア言語についての展開を行う。

b. JSTでは、科学技術文献検索・翻訳の試行的サービスの検証・評価結果を踏まえ、下記の自立的な取り組みについて検討する。

・翻訳システムの実用システムを利用して中国語科学技術文献の日本語への翻訳に活用することにより、中国科学技術情報の日本国内における流通の促進を目指す。

・翻訳システムの実用システムを利用してＪＳＴが保有する日本語の科学技術文献データベースに対して中国語を付与し、アジア地域に向けての情報発信の促進を目指す。
・研究成果で得られた半自動辞書構築システムを用いて、科学技術用語辞書の充実を図ることにより、中国およびアジア多言語の情報検索用辞書の構築の推進に寄与する。また、本研究により開発した言語資源・技術を企業に移転することにより、企業の競争力を増強し、市場において新たな需要を生み出す手助けをする。

6. 生命倫理・安全面への配慮について

特に無し。

7. 具体的な達成目標（ミッションステートメント）

a. 3年の中間段階におけるステートメント

○対訳コーパス（100 万文規模）の作成（データの収集、翻訳）と半自動解析

○辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の作成

○辞書半自動構築システムの構築と評価

○解析・翻訳エンジンの開発･改良

○特定の対象分野について、日中機械翻訳プロトタイプ実証システムの動作を確認し、情報通信研究機構のサイトなどで日中機械翻訳を体験できるようにする。

b. 研究終了段階におけるステートメント

○日本語および中国語の科学技術文献を対象に、翻訳率80％以上を実現する高品質な日中・中日機械翻訳プロトタイプシステムを開発する。

○科学技術文献検索・翻訳の試行的なサービスを行ない、その有用性を検証・評価する。

○実施期間中随時、及び終了後速やかに、著作権をクリアした上で、他者が使用できる形態で、成果となる対訳コーパス、翻訳用辞書、翻訳エンジン、および評価に使用したデータセットを公開する。

8. 実施体制について

氏名	所属機関・職名	提案課題における役割
◎　井佐原　均	（独）情報通信研究機構・上席研究員	研究代表者
○　黒橋　禎夫	京都大学・教授	サブテーマ1責任者
内山　将夫	（独）情報通信研究機構・主任研究員	サブテーマ1、サブテーマ3参画者
○　辻井　潤一	東京大学・教授	サブテーマ2責任者
中川　裕志	東京大学・教授	サブテーマ2参画者
梶　博行	静岡大学・教授	サブテーマ2参画者
菊池　俊一	（独）科学技術振興機構・次長	サブテーマ2参画者
○　内元　清貴	（独）情報通信研究機構・主任研究員	サブテーマ3責任者、サブテーマ1参画者

9. 各年度の計画と実績

a. 平成 18 年度

・計画

・対訳コーパス（100 万文規模）の~~完成~~作成

・対訳コーパス（100 万文規模）半自動解析システム、解析・翻訳エンジンの基本・概念設計

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）のデータ収集と作成着手

・実績

・対訳コーパス（100 万文規模＝2500 万文字規模）の構築中。このうち、約 1/3 が完成。

・対訳コーパス半自動解析システム、解析・翻訳エンジンの基本・概念設計、基本詳細設計を終了。

・対訳コーパス自動解析システム、解析・翻訳エンジンの基本実装、動作確認を終了。

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）作成のためのデータ収集を完了。

・日英中専門用語辞書作成手法の基本実装と動作確認を終了。

・辞書自動構築システムのための用語抽出手法の基本実装と動作確認を終了。

・辞書自動構築システムのための曖昧性解消手法の基本実装と動作確認を終了。

b. 平成 19 年度

・計画

・対訳コーパス（100 万文規模）の作成

・対訳コーパス（100 万文規模）の半自動解析システムの詳細設計および第 1 版の完成

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の作成

・辞書自動構築システム、解析・翻訳エンジンの基本詳細設計

・実績

・対訳コーパス（100 万文規模＝2500 万文字規模）の構築中。このうち、約 2/3 が完成。

・対訳コーパス半自動解析システムの詳細設計と第 1 版の実装を完了。

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）を試作。

・辞書自動構築システム、解析・翻訳エンジンの基本詳細設計を終了。

c. 平成 20 年度

・計画

・対訳コーパス（100 万文規模）の完成

・対訳コーパス（100 万文規模）半自動解析システムの評価

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の第 1 版完成。

・辞書半自動構築システムの詳細設計および第 1 版完成

・解析・翻訳エンジンの実装

・日中機械翻訳プロトタイプシステムの動作確認、試行的な公開。

d. 平成 21 年度

・計画

・対訳コーパスの評価および拡充

・対訳コーパス（100 万文規模）半自動解析システムの改良と評価

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の第 1 版の評価

・辞書半自動構築システムの評価

・解析・翻訳エンジンの改良および完成

・日中・中日機械翻訳プロトタイプシステムの改良

・解析・翻訳エンジンの評価手法についての概念設計

e. 平成 22 年度

・計画

・対訳コーパス（100 万文規模）半自動解析システムの応用実験

・辞書半自動構築システムの運用開始

・辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の第 2 版の作成

・解析・翻訳エンジンの評価

・日中・中日機械翻訳プロトタイプシステムの完成および実証実験

・科学技術文献検索・翻訳の試行的なサービスの実施と検証

10. 年次計画

研究項目	18 年度	19 年度	20 年度	21 年度	22 年度
(1) 日中・中日の用例ベース翻訳のための要素技術の研究開発 (参画研究機関) (独) 情報通信研究機構 京都大学	中国語の形態素解析・構文解析の高度化 32（百万円）	14（百万円）	9（百万円）	30（百万円）	
	用例ベース翻訳における用例利用の柔軟化 12（百万円）	16（百万円）	12（百万円）	17（百万円）	20（百万円）
	対訳コーパスの語・句の高精度対応付け 25（百万円）	20（百万円）	20（百万円）	5（百万円）	
(2) 日中・中日言語資源の構築と構築技術に関する研究開発 (参画研究機関) 東京大学 静岡大学 (独) 科学技術振興機構	日中・中日翻訳用大規模辞書の作成手法の開発および、専門用語の抽出と意味関係の認定 27（百万円）	22（百万円）	31（百万円）	36（百万円）	42（百万円）
	訳語選択のための語義ネットワークの生成方法 12（百万円）	12（百万円）	12（百万円）	12（百万円）	15（百万円）
	大規模日中パラレルコーパスの構築 28（百万円）	34（百万円）	34（百万円）	言語資源の評価・拡充 23（百万円）	12（百万円）
		半自動辞書構築システムの開発 0（百万円）	2（百万円）	システムの評価・運用 4（百万円）	4（百万円）
(3) 日中・中日機械翻訳プロトタイプシステムの開発および実証実験 (参画研究機関) (独) 情報通信研究機構		日中・中日機械翻訳プロトタイプシステムの作成と評価 17（百万円）	30（百万円）	22（百万円）	56（百万円）
(4) 研究運営委員会 (参画機関全機関)		アウトリーチ活動（シンポジウム開催等） 3（百万円）	4（百万円）	4（百万円）	4（百万円）
	4（百万円）				

11. 研究運営委員会

委員	所属	備考
(研究実施者) ○ 井佐原　均	情報通信研究機構　けいはんな研究所　上席研究員	代表者
(外部有識者) 田原　康夫 勝野　頼彦 田中　穂積 松本　裕治	 総務省　情報通信政策局　技術政策課　研究推進室　室長 文部科学省　研究振興局　情報課　課長 中京大学　大学院情報科学研究科　教授 奈良先端科学技術大学院大学　情報科学研究科　教授	