科学技術文献を対象とする日中機械翻訳システム開発プロジェクト

井佐原均（NICT), 黒橋禎夫（京大), 辻井潤一（東大), 内元清貴（NICT),
中川裕志（東大), 梶博行（静岡大), 中村徹（JST)

1. はじめに

科学技術が発展し、世界中が瞬時に情報を共有できるようになった現在、世の中に存在するさまざまな知識を活用できるかどうかが、個人の生活の充実に大きな影響を与えるようになっています。知識の多くは言葉によって表現されており、日本語・英語・中国語といった言語の違いが、知識の流通と利用の大きな妨げになっています。このような言語障壁の克服をめざし、人間の言葉をコンピュータで処理する自然言語処理技術の高性能化が進められてきました。

今回、情報通信研究機構（NICT)、科学技術振興機構（JST)、京都大学、東京大学、静岡大学の５機関は協力して、今年度から５年間で科学技術文献を対象とする日中・中日機械翻訳システムを開発します。この研究開発の一部は、科学技術振興調整費・重要課題解決型研究等の推進「日中・中日言語処理技術の開発研究」として実施されます。主として解析・翻訳エンジンの開発、プロトタイプシステムの開発・改良を情報通信研究機構および京都大学が担当し、言語資源の収集、構築を東京大学、静岡大学、科学技術振興機構が担当します。

このプロジェクトでは、５年間の開発期間で、日中の科学技術文献を対象とした、実用的な機械翻訳システムを開発します。翻訳手法としては、言語の構造をより深く考慮した用例ベース翻訳を用います。この手法の実現のためには、大量の用例（対訳コーパス）を蓄積する必要がありますが、我々は１千万文規模の日中対訳コーパスを開発する予定です。また、科学技術文献の翻訳・情報検索用辞書を対訳コーパスから半自動作成する手法を確立します。今回は特に、アジア諸国の言語のうち中国語に焦点をあて、大規模な言語資源（日中対訳コーパス、日中・中日辞書）を構築し、その言語資源を用いた機械翻訳プロトタイプシステムを開発して実証実験ならびに評価を行ないます。

基盤となる技術として、日本語や中国語の解析システムの性能向上を図ります。また、用例翻訳の手法を科学技術文献等の長く複雑な文にも対応できるように改良を進めます。

プロジェクトの途中段階でも、コーパス等の言語資源は可能な限り研究用に公開する予定です。また、アウトリーチ活動として、研究の内容や成果を出来るだけわかりやすく、広く発信していくことを目指します。今回、開発するシステムの概念図を図１に示します。実際のシステムは日中・中日双方向のシステムです。

2. 中国、そしてアジアへ

欧米諸国と比べて、特にアジアにおいては英語での情報流通には困難が伴います。今回の研究開発においては、アジア諸国の一員としてのわが国の責務として、アジア言語の機械翻訳の実現を目指し、その第一歩として、特に科学技術の進展が顕著である中国を対象に科学技術文献を主たる対象とする機械翻訳システムの開発を行うこととしました。

この実現により、言語障壁のために中国国内のみで流通している有益な科学技術情報を、我が国の研究者・技術者、事業者が容易に活用することができるようになり、共同研究事業の設立など大きなビジネスチャンスにも繋がるでしょう。また、日本が最先端を担う科学技術分野の文献が中国国内で流通することにより、中国における科学技術の発展も期待できます。

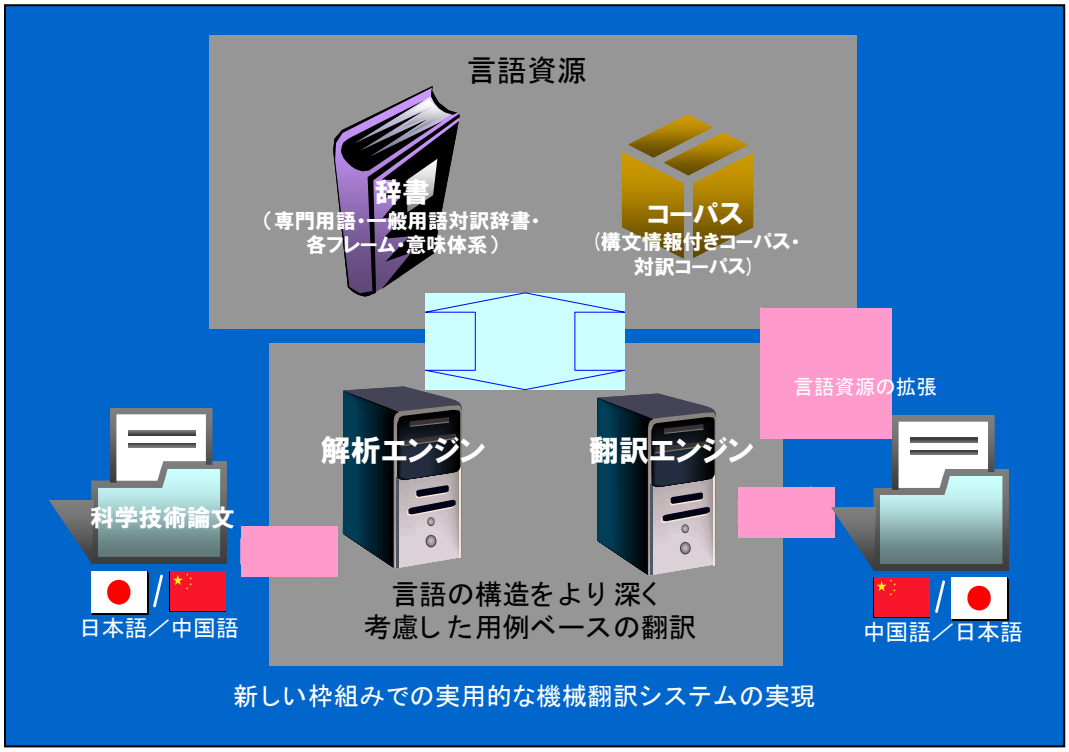

図１　日中・中日翻訳システムの概要

さらに、他のアジア言語の言語資源を整備すれば、解析・翻訳エンジンを大幅に変更することなく、それらの言語への展開が可能となります。言語資源の整備も本研究の成果の利用により効率良く行なえるようになることが期待できますので、日本語や中国語以外のアジア言語についても比較的容易に母国語で科学技術情報を検索・閲覧できるようになり、アジア各国の科学技術文献情報がアジア地域全体で流通しやすくなると期待されます。今回のプロジェクトの参画機関の一つである NICT では、これまで、中国、インド、東南アジア諸国との研究交流を進めており、特にタイには、タイ自然言語ラボラトリー（TCL）を設置しています。将来的には TCL を介して、今回の機械翻訳システムの技術を広くアジア言語を対象に広げることを検討しています。

3. 手法(統語情報を用いた用例翻訳)

これまで、機械翻訳の実装には、変換方式や中間言語方式などが用いられてきました。いずれの方式も人手によるさまざまな知識（文法規則、単語辞書、意味辞書など）の作成が必要であり、そのような知識の一貫性を持った作成は非常に困難です。

一方、人間が翻訳をするときには、このような知識を適用しているのではなく、過去に読み聞きしたことのある類似した文の訳を組み合わせて、翻訳をしているだろうという考え方から、用例翻訳の考えが１９８１年に提案されました。当時はコンピュータの能力が十分ではなく、この手法を実用的なシステムに実現することは出来なかったのですが、近年、コンピュータの能力が向上し、また、文をそのまま例として使うのではなく、文法的に解析した上で、入力文と蓄積された用例文の類似性を判定する手法が開発されたことにより、実用的な用例翻訳システムを開発する基盤が整ってきました。

用例翻訳では、膨大な対訳コーパスに含まれる例文と入力文の類似性を利用して翻訳します。ここで必要となるのは用例となる対訳コー

パスと、文同士の類似性を判定する方法であり、大規模な規則を作成する必要はありません。また、用例を追加することによって翻訳の質が向上する、また、用例の訳には前後関係による訳の違いが自然に含まれており、機械翻訳の訳文にもそれが反映される、といった特徴があります。

4. 研究開発項目

今回の研究開発のうち、科学技術振興調整費に関わる部分では、以下の項目の研究開発を行います。これらの関係を図２に示します。

① 日中、中日の用例ベースの翻訳システムの研究開発

○ 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ります。

○ 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します。

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

○日中・中日翻訳用大規模辞書の構築

日英(及び英中)の専門用語と一般用語の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（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します。

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します。

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因果関係などの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します。

研究開発用に、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、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します。

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します。

③ 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

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します。

5. ミッションステートメント

今回の研究開発のうち、科学技術振興調整費に関わる部分では、以下のミッションステートメントに沿った開発を行います。

平成２１年度末の中間段階では、科学技術分野の大規模日中・中日辞書の作成を終え、情報通信研究機構のサイトなどで特定の分野を対象とした日中機械翻訳を体験できるようにします。平成２３年度末の課題終了までには、１００万文規模の日中対訳コーパスを構築し、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します。また、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・評価します。成果は随時公開します。

6. おわりに

科学技術の目的は、あまねく人々に能力や地位に関わらず平等にすばらしい生活を提供することにあります。私たちはコンピュータに言語を処理する能力を与えることにより、世界中の人々が能力（語学力）や地位（通訳を雇えるか）に関わらず、言語障壁を意識せずにすむ環境の実現を目指しています。今回の機械翻訳システムの研究開発により、そのような目標の一端が達成できると期待しています。