4　多言語翻訳技術

4 Multi-Lingual Translation Technology

4-1　多言語高精度を実現する専用翻訳システム

4-1 Special-Purpose System for Multi-Lingual High-Quality Translation

隅田英一郎
SUMITA Eiichiro

要旨

NICT は専門分野を限定しつつ、高精度自動翻訳システムも実現するための翻訳技術を研究開発している。音声翻訳では旅行会話に注力し、テキスト翻訳では e コマースの説明文に注力し、事業化に至る成果をあげてきた。本稿では、同技術の概要を述べる。

NICT is conducting research for realizing high-quality automatic translation system while restricting the domain of translation. We've been concentrated on travel conversation in speech translation and explanation of products in text translation, and recently we put our technology on a commercial basis. In this paper, we outline the technology.

[キーワード]
自動翻訳，音声翻訳，TEXT 翻訳，コーパスベース翻訳
Automatic translation, Speech translation, Text translation, Corpus-based translation

1　高精度の自動翻訳

NICT は専門分野を限定しつつ、高精度の自動翻訳システムも実現するための翻訳技術を研究開発している。一方、一般には、汎用の翻訳システムを構築することを目指した研究開発が従来より行われてきた。例え話でいえば、前者は刺身包丁を作ることであり、後者は万能包丁を作ることである。後者の技術でできた包丁は魚でも肉でも野菜でも何でも切れるが、その切れ味は鈍く、生魚の刺身は引き千切った様な別物になってしまう。同様に、現在利用可能な日英翻訳システムは汎用だが、その翻訳品質はよろしくなく。このために自動翻訳システムは役に立たないという印象を持つ人が多くなってしまっている。

音声翻訳（『MASTAR プロジェクトにおける音声翻訳技術』、本特集号 **7-1**）では旅行会話に注力し、テキスト翻訳では e コマースの説明文に注力し事業化に至るという成果をあげてきた。本節では、同技術の概要を述べ、関係する節へのリンクを示す。

2　多言語の自動翻訳

言語は人類の最大の壁の 1 つである。自動翻訳はこの壁を超える究極の手段として期待されている。

例えば、検索エンジンの普及で、我々は日本に居ながらにして、世界中の情報に簡単にアクセスできる。しかし、この情報が外国語で表現されている場合、多くの日本人にとっては暗号と同じで活用できる人は少ない。インターネットでの言語使用の状況を調べると、上位 10 位までの言語（英語、中国語、スペイン語、日本語、フランス語、ドイツ語、ポルトガル語、アラビア語、韓国語、イタリア語）で、84％のシェアである（図 1）。

日本語だけだと7%に過ぎない。日本語以外の9言語から日本語への高精度の自動翻訳システムが作れれば、インターネット上の情報の84%が分かるということになり、日本人の情報の受信能力を飛躍的に高めることになる。逆に、日本語から日本語以外の9言語への高精度の自動翻訳システムが作れれば、日本人の情報の発信能力を飛躍的に高めることになる。

そこで、実際にどうしたらよいかと考えてみると、これらの10の言語は、文字、単語、文法など様々な面で大きく異なるので、言語特性にあまり依存せず高品質を実現する自動翻訳技術が必要になる。

3 コーパスベース翻訳技術

ここでは、**2**で述べた課題を解くための手法、即ち、多言語高精度の自動翻訳システムを実現するためのコーパスベース翻訳技術について述べる。

コーパスベース翻訳技術とは、対訳コーパス（同じ意味の原文と訳文の対を集めたもの）から、翻訳システムの知識（確率付きの対訳辞書等の翻訳に必要な知識）を自動的に構築する（図2）技術である。この自動構築に由来する2つの利点がある。（A）新しい分野の翻訳システムを作るには、その分野の対訳コーパスを集めることができれば高精度を達成できる可能性がある。例えば、新聞、特許、マニュアル、自治体の発信する情報、医療や介護にかかわる情報、WEB、ブログ等、どんな分野でも、その分野の対訳データを集めれば、専用の翻訳システムが構築でき、高精度を達成できる可能性がある。実際にeコマースの説明文に注力し事業化に至るという成果をあげてきたし、前記の旅行会話もこのような専門分野の一例である。（B）N個の言語からなる多言語対訳コーパスを用意すれば、全ての組合せである$N(N\text{-}1)$個の翻訳システムが自動的に構築できること。我々は、既に旅行会話の分野で多言語対訳コーパス（$N = 21$）を構築し、全ての組み合わせ（420通り）の翻訳システムを実現し、それらが十分に実用レベルの翻訳品質を達成していることを確認し、VoiceTra/TexTraというiPhoneアプリケーションとして公開している。

4 研究の2つの柱

コーパスベース翻訳技術で高精度の自動翻訳を実現するためには、大きく2つの研究課題がある。

①**対訳データ収集**：ある一定量以上の対訳データが集まると翻訳品質が実用レベルになることがわかっており、対訳データを経済的に短期間で収集する手法を確立することが重要である。

②**翻訳アルゴリズムの高度化**：同じデータ量でもアルゴリズムによる性能差が大きいことがわかっており、良いアルゴリズムの研究

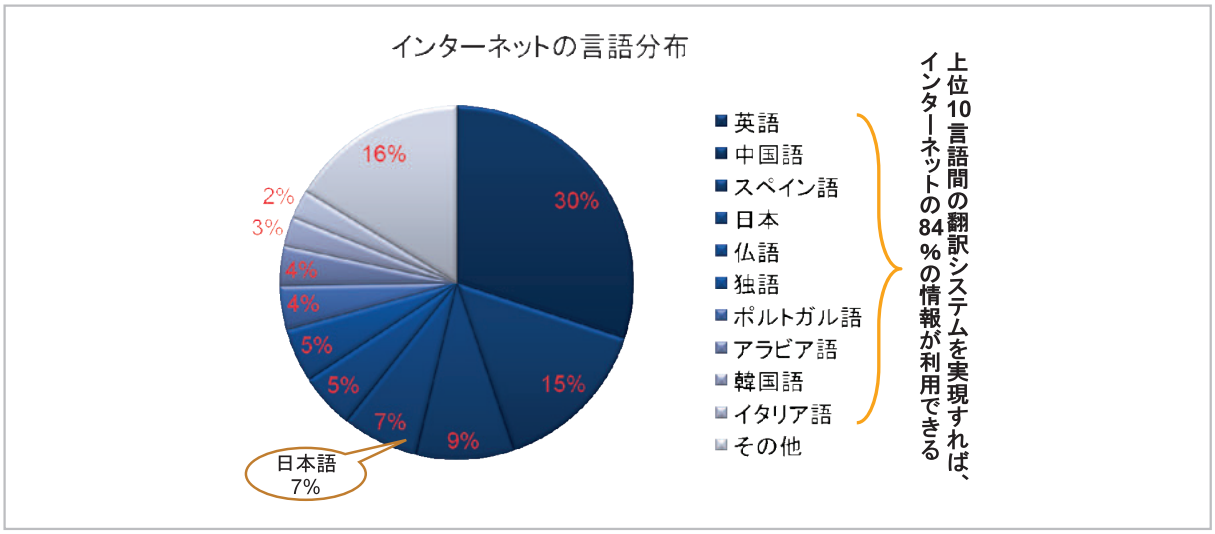

図1 インターネット上の多言語情報の割合

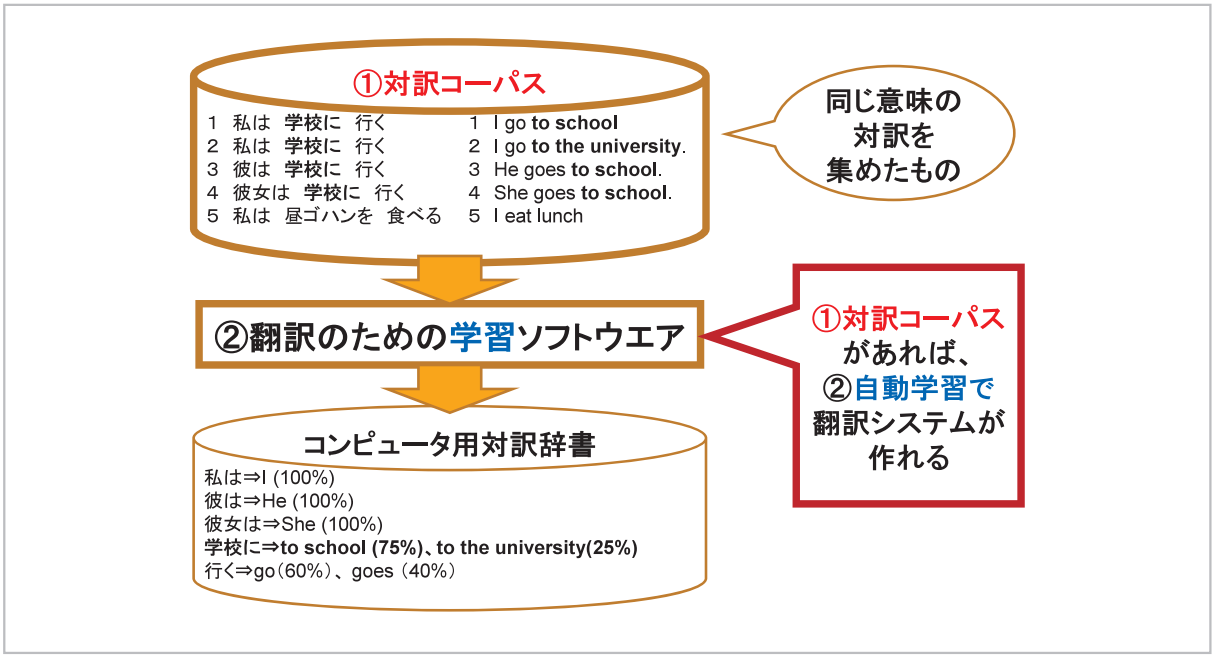

図２ コーパスベース翻訳技術の基本

が重要である。

以下で、順に例をご紹介する。

4.1 対訳データ収集

コーパスベース翻訳技術の主たる知識源は対訳データであり、これを効率的に収集することが重要である。そこで、2つの補完的なアプローチを並行して進めている。(1) WEB クローリング、単言語コーパスからの対訳創出、2 言語類似コーパスの利用などのコンピュータ中心のアプローチ。(2) WEB に散在する対訳の収集、ボランティア翻訳のホスティング・サービス、外部機関との提携など、人や社会中心のアプローチ。詳細は、『対訳データの効率的な構築方法』（本特集号 **4-2**）を参照。

4.2 翻訳アルゴリズムの高度化

翻訳アルゴリズム高度化にも、多くのサブテーマがある。語分割の高精度化、単語対応プログラムの高精度化、固有名詞処理、翻字処理（『ベイジアンアライメントに基づく翻字システムと機械翻訳への応用』本特集号 **4-3**）や専門用語の自動獲得、分野や話題への適応、構文の導入、場面・状況・文脈のモデル化、複数翻訳を最適に混合する手法、モデル学習の並列化、など。

ここでは、「構文の導入」について説明する。日本語と韓国語、スペイン語とイタリア語のように互いに似た言語間では問題になりにくいが、日本語と英語のように互いに似ていない言語間では語順が問題となる。日本語の基本の語順が SOV であり、英語のそれが SVO であり、このような場合に正しい語順で訳文を生成することが困難な課題になる。我々は、全ての語順の可能性を素朴に許すのではなく、入力構文で制約して、条件に合うものだけ計算する手法を提案している。これにより、翻訳仮説数の大幅な削減を実現し、日英間の翻訳の誤り率を低減できることを確認した。また、多重翻訳仮説の融合に構文を利用する手法については、『構文情報を直接利用した機械翻訳システムコンビネーション』（本特集号 **4-4**）を参照。

5 高精度専門翻訳の事例

5.1 電子通販

高精度翻訳システムが求められる分野として、電子通販（e コマースとも呼ぶ）がある。電子通販は成長産業であり且つ海外進出が課題となっており、膨大な商品の量、商品回転の速さから自動化が必須だが高品質システムが存在しなかった。NICT の①翻訳支援技術による対訳の効率的構

築、②対訳辞書自動構築技術による専門用語辞書の効率的構築、③構文に基づく統計翻訳技術を組み合わせて電子通販向け高精度翻訳システムを実現して事業者に技術移転し、国内最大級アパレル電子通販のグローバルサイトで活用されている。

5.2 特許の翻訳

特許庁の国際知財戦略（Global IP Initiative）～国際的な知的財産のインフラ整備に向けた具体的方策～ 2011 年 7 月*などにあるように、中韓特許文献が増大し、係争案件も増加している現状を踏まえ、中・韓→日への翻訳機能を備えた外国特許文献検索システムの整備を行うことが国民の利益になる。

文長が長くなる特許文を対象として、長文翻訳のための新しい技術を研究開発している。①文分割法：長文を表層の特徴によって分割し翻訳結果を統合する手法と②名詞句カプセル化法：名詞句をカプセル化し、文を短縮して翻訳、名詞句の翻訳を埋め戻す手法を創出し、これらを併用して、大幅な性能改善を実現した（図 3）。

また、NTCIR9（2010 ～ 2011）の中で、特許対訳コーパスを提供し翻訳性能を比較するコンペ型国際会議 PatentMT を NII と共催した。米欧アから IBM や BBN を含む 21 研究機関を集結し、アルゴリズムの進展で英日、中英で統計翻訳が規則翻訳より有望であることを明らかにした。

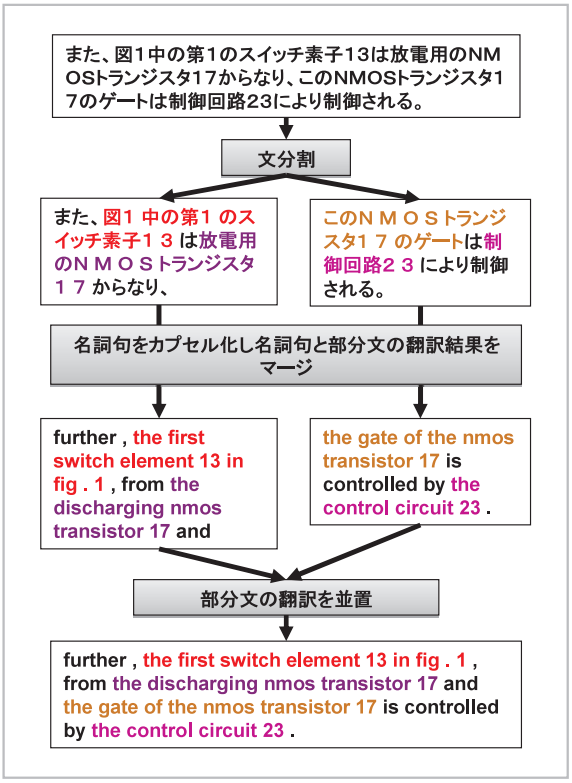

図 3 長文翻訳の事例

6 今後の進め方

今後、どんな言語でもどんな分野でも翻訳できるようにするために 3 つのステップを考えている。

ステップ 1 として、コーパス構築の手法・基盤の確立、翻訳アルゴリズムの高度化、いくつかの分野での翻訳システムの「モデル開発」を行う。

ステップ 2 として、多分野・多言語コーパスを実現するための社会経済的に「回る」仕組を提案する。

ステップ 3 として、言語翻訳技術の「見える化」を進め、どんな分野でもどんな言語でも翻訳できるように外部機関を巻き込んだ活動を進める。

＊ http://www.jpo.go.jp/shiryou/toushin/shingikai/pdf/tizai_bukai_16_paper/siryou_01.pdf

参考文献

1 隅田 英一郎, “MASTAR プロジェクトにおける音声翻訳技術,” 情報通信研究機構季報, 本特集号, 7-1, 2012.
2 内山 将夫, “対訳データの効率的な構築方法,” 情報通信研究機構季報, 本特集号, 4-2, 2012.
3 Finch Andrew, 安田 圭志, “ベイジアンアライメントに基づく翻字システムと機械翻訳への応用,” 情報通信研究機構季報, 本特集号, 4-3, 2012.
4 渡辺 太郎, “構文情報を直接利用した機械翻訳システムコンビネーション,” 情報通信研究機構季報, 本特集号, 4-4, 2012.

（平成 24 年 6 月 14 日 採録）

隅田英一郎（すみた えいいちろう）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室室長
博士（工学）
自然言語処理、機械翻訳
eiichiro.sumita@nict.go.jp