# 多言語自動翻訳技術

## ー世界の多様な言語を相互に翻訳するための技術開発をめざしてー


**隅田 英一郎**（すみた えいいちろう）

**知識創成コミュニケーション研究センター 言語翻訳グループ グループリーダー**

1982年電気通信大学大学院修士課程修了。1999年京都大学博士（工学）。
現在、NICT言語翻訳グループ グループリーダー、神戸大学大学院システム情報学研究科客員教授。
機械翻訳、eラーニングを研究。


### はじめに

インターネットでの言語使用の状況は、上位10位までの言語で、84%のシェアになります。日本語は第4位で7%に過ぎません。日本語以外の9言語から日本語へ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%が読めるようになり、日本人の情報の受信能力を10倍以上高められます。発信も同様です。10言語の間の自動翻訳システムはどうしたら実現できるでしょうか。各言語は、文字、単語、文法など様々な面で他の言語と異なりますので、個別言語の特性に依存せず実現できる自動翻訳技術が必要になります。

### 統計翻訳技術による多言語翻訳

ハードウエアの処理速度や記憶容量が格段に進歩したこと、文章や辞書が大量に計算機上に集積されるようになったこと、などを受けて、自動翻訳の研究において、対訳コーパス（同じ意味の原文と訳文の文レベルの対を集めたもの）から、翻訳に必要な知識を自動的に構築する技術が興り、現在、主流の研究パラダイムとなっています。例えば、統計翻訳と呼ばれる技術（図1）では、対訳コーパスから2言語間の対応関係をモデル化する翻訳モデル（直感的にいうと、確率付き対訳辞書です）と目的言語らしさをモデル化する言語モデル（例えば、英日翻訳の場合、日本語の単語の並びの自然さを表す確率付き日本語辞書です）を導出し、両者に基づく確率を最大化するように翻訳します。*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*ー1）個の翻訳システムが自動的に構築できます。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、全ての組み合せである420通りの翻訳システムを実現し、実用レベルの翻訳品質（図3）を達成しています。

### 統計翻訳高度化の2つのポイント

さて、その統計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つの研究課題があります。①ある一定量以上の対訳コーパス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いますので、対訳コーパス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になります。また、②同じデータ量で


図1●統計翻訳技術の概要


図2●多言語翻訳の出力画面（日本語から多言語への翻訳で、ベトナム語が選択されているところ）

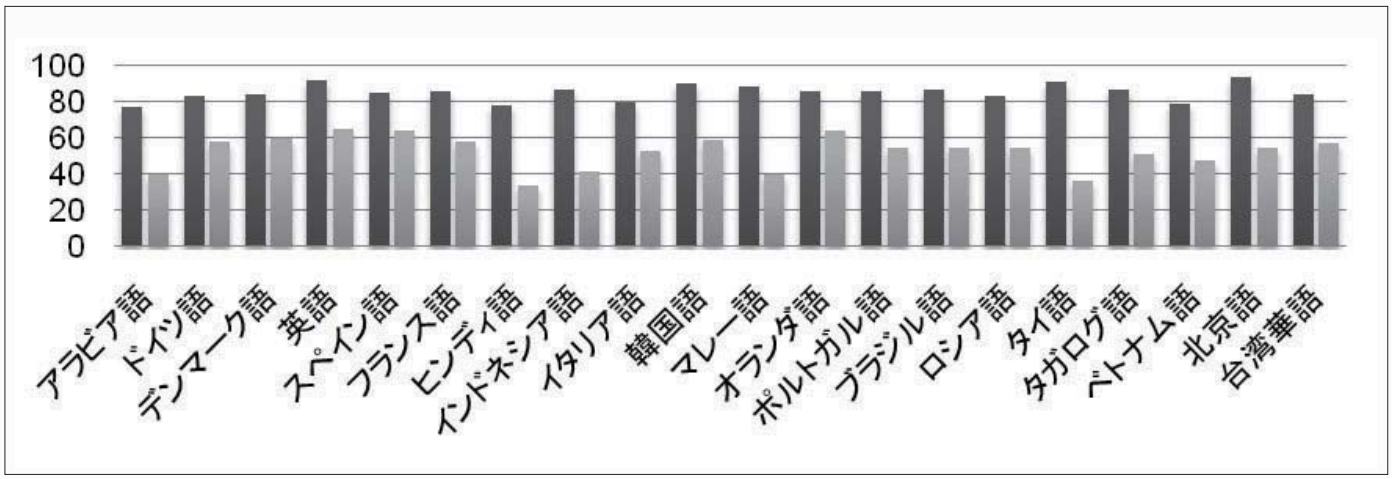


図3●翻訳率の比較（広く利用されているソフトウェア（淡色）とNICTのソフトウェア（濃色）と比較。 縦軸が日本語への翻訳率、横軸が翻訳元の言語）

もアルゴリズムによる性能差が大きいことがわかっていますので、与えられたデータで高精度を実現する良いアルゴリズムの研究が重要になります。以下、順にご紹介します。

## 対訳コーパス収集

対訳コーパスを効率的に収集するために、2つの補完的なアプローチがあります。（A）WEB から対訳コーパスをクローリングすることや文章レベルの対訳から自動的に文レベルで対応付けする技術などのコンピュータ中心のアプローチと（B）ボランティア翻訳のホスティング・サービス[*1]や外部機関との提携など、人や社会中心のアプローチです。NICT の言語翻訳グループでは、両方のアプローチを併用して精力的に対訳コーパスを集めています。例えば、自動文対応技術で、新聞やマニュアルなど様々な分野の対訳を集めています。特に、特許に関しては1,800万文の日英対訳コーパスを構築しました。これは現在公開されているどの対訳コーパスよりも大きい世界最大規模です。NICT はこれらの有用なデータを我が国の企業や大学に高度言語情報融合フォーラム[*2]を通じて公開を開始しています。

## 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ります。日本語や中国語などで必要となる単語分割の高精度化、大量の固有名詞等を音に従って翻訳する翻字処理（New York をニューヨークと変換すること）や、複数の翻訳を最適に混合する手法、など。ここでは、単語分割について説明します。多言語翻訳を効率的に実現する目的で、各言語の単語分割プログラムの現状を考えると母国語話者による研究が遅れていたり、種々の条件から、入手困難な場合もあり、一様ではありません。また、既存のプログラムが翻訳に最適と限りません。NICT はこの状況を考慮して、分割の初期値として文字を設定し、翻訳スコアが上昇するように単位を大きくする手法を提案し、多言語で検証しました。表1にあるようなアラビア語、タイ語、ベトナム語をはじめ、翻訳率は改善でき、言語によっては既存の単語分割プログラムより高い翻訳率を得ることができました。

表1●多様な文字の言語でも高い翻訳品質を実現するための多言語向け単語分割法

| 言語 | サンプル | Baseline | 提案法 |
|---|---|---|---|
| アラビア | نعم ، انه كذلك. | 58.60 | 63.70 |
| タイ | ใช่มันเป็นเช่นนั้น | 44.41 | 55.00 |
| ベトナム | Vâng, đúng rồi. | 49.91 | 60.56 |

## おわりに

現在、専門分野向けの多言語の高精度翻訳技術の可能性を実証したところですが、今後は、まず、多分野化を進めるとともに、全く新たな分野へ自動翻訳技術を適用して、実用性を証明していきたいと考えます。

また、中国語、韓国語などアジア言語に注力し、アジア諸国との情報の受発信に貢献し、成長するアジアと日本の連携に役立っていきたいと考えます。

さらに、NICTの翻訳は、機械と人間の協調が特徴であり、強みでありますから、この面をさらに強化していきます。

### 参考情報

*1 内山将夫、「みんなの翻訳」NICT NEWS 2009年6月号
http://www.nict.go.jp/publication/NICT-News/0906/04.html

*2 高度言語情報融合フォーラム
http://www.alagin.jp/

*著者及び NICTの承諾を得て NICTニュース 2011年3月号の記事を掲載しています。