身近になった多言語自動翻訳

ー半世紀以上の研究開発を経て実用化されつつある技術ー

内山 将夫（うちやま まさお）
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 主任研究員

自動翻訳の研究を 10 年くらいしています。自動翻訳の性能をあげることにより、世の中が住みやすくなることに貢献したいと思っています。家族にわかる研究をするように努めています。

「自動翻訳は、長年の研究開発を経て実用化されつつあります。NICT の自動翻訳エンジンは、対訳データから自動構築可能です。この研究成果を社会に還元したいと思っています。」

身近になった多言語自動翻訳

自動翻訳の研究は、1940 年代から始まりました。日本でも、1980 年代に産学官で盛んに研究開発されて、たくさんの商用自動翻訳システムが開発されました。そして、現在では、インターネットのポータルサイトなどで、自動翻訳が無料で提供されていますし、商用の自動翻訳システムの提供も盛んです。

自動翻訳システムの種類には、基本的には、人手で記述した規則に基づいたシステムと、大量の対訳テキストから自動的に翻訳規則を学習するコーパスベースの自動翻訳システムの 2 種類があります。コーパスというのは、大量のテキストからなるデータベースのことです。

これまでは、一般的に利用されている自動翻訳システムは、人手で記述した規則に基づくシステムでしたが、最近は、コーパスベースの自動翻訳システムの性能も向上しています。NICT が研究している自動翻訳システムは、コーパスベースの自動翻訳システムです。その研究開発の基盤には、大きく分けると、ユーザ、言語資源、アルゴリズムの 3 点があります。

ユーザ

自動翻訳システムの目的は、ユーザの役に立つことです。そのため、ユーザは、自動翻訳システムの研究において、もっとも尊重する必要があります。

NICT では、旅行会話専用の自動音声翻訳システムとして VoiceTra を開発しています (p.140-143 参照)。また、e コマース用の自動翻訳エンジンとして、日本最大級のアパレルサイト

図1　みんなの翻訳　(http://trans-aid.jp)

の、韓国サイトについて、日韓自動翻訳により商品説明文を韓国語に訳すサービスを提供しています。この自動翻訳エンジンのための対訳データを作成するために、株式会社バオバブとの共同研究により「留学生ネットワーク@みんなの翻訳」(https://en.ecom.trans-aid.jp/) を開発し、留学生のアルバイトにより効率的に対訳データを作成しました。また、ボランティアによる人手翻訳を支援するために東京大学図書館情報学研究室と共同で「みんなの翻訳」(http://trans-aid.jp) を運営しています(図 1)。

このように、NICT では、最新の研究成果を一般に利用していただくことにより、研究成果を社会に還元すると同時に、そのフィードバックを研究開発に役立てています。

言語資源

最も重要な言語資源は、対訳テキストです。統計的自動翻訳では、分野を限定した場合で十数万文、分野を限定しない場合には 1,000 万文以上の対訳文が翻訳エンジンの訓練に必要です。図 2 は、e コマース分野における、訓練に利用した対訳文数と翻訳精度の関係を示しています。なお、A と B は商用の翻訳エンジンですが、これらは対訳文での訓練はしていないため、一定の精度です。

NICT では、異なる言語の文書から自動的に文と文の対訳を作成する技術を開発しています。そして、この技術を利用することにより、日本と米国に同時出願された特許文書から 1,000 万文以上の大規模な日英対訳コーパスを自動作成しました。今後は、このコーパスを利用して、日英の特許翻訳や類似文検索などのサービスを開発する予定です。

また、NICT では、新聞記事から 25 万文規模の日英対訳コーパスも作成しており、この対訳コーパスは、ライセンス契約により第 3 者も利用可能です。このコーパスは英辞郎に採用予定ですので、たとえば、「英辞郎 on the Web」で NICT が提供した対訳文が検索可能となります。

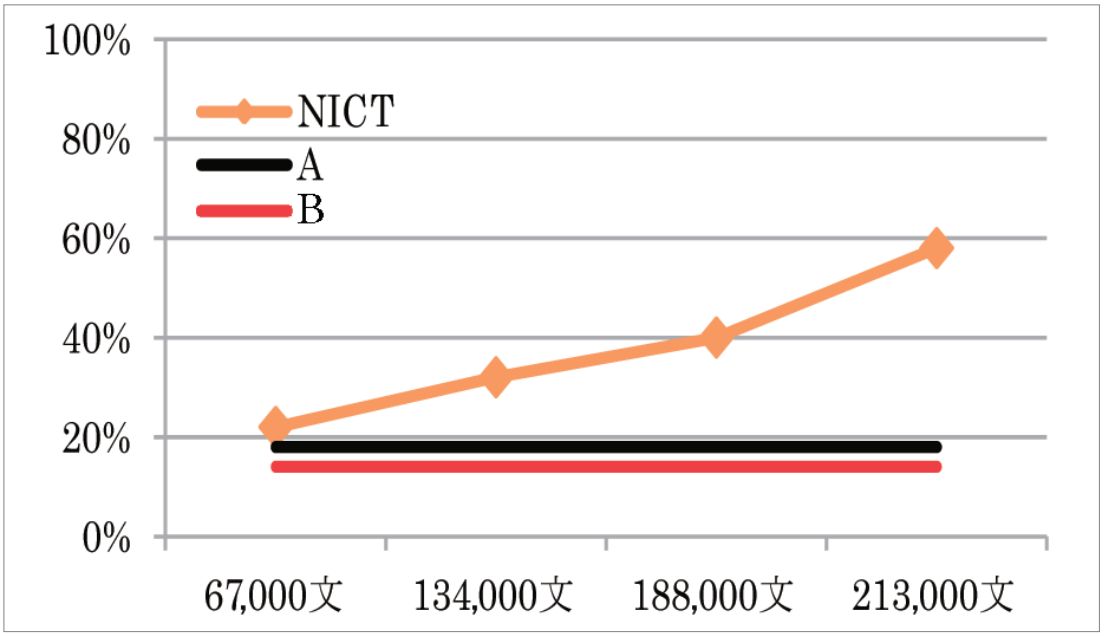

図2　日英翻訳の性能改善

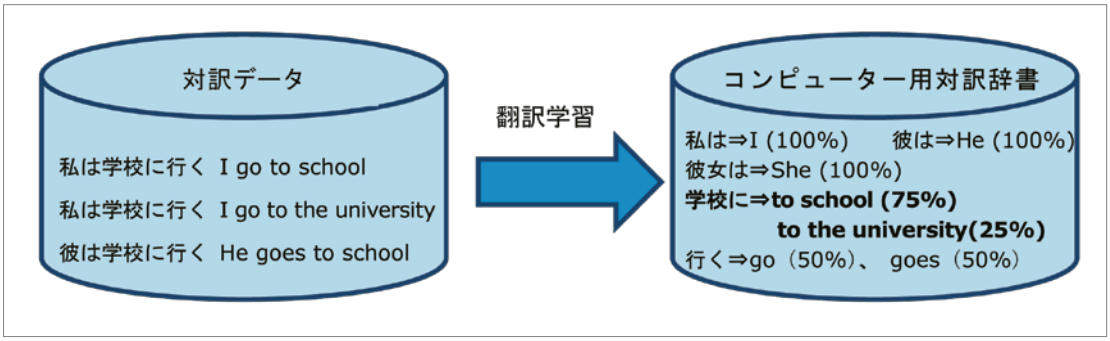

図3　コンピューター用対訳辞書の作成

アルゴリズム

コーパスベースの自動翻訳は、1980 年代に、長尾真氏により提案されました。その後、1990 年代に、IBM により、対訳コーパスから自動的に対訳辞書や翻訳規則を推定するアルゴリズムが開発されました。2000 年代には、自動翻訳を(不完全ながらも)自動的に評価する方法が開発されました。現在では、英日や日英などの言語構造が離れた言語対の翻訳も、ある程度の精度で自動翻訳ができるようになりました。

コーパスベースの自動翻訳では、対訳コーパスから自動的に翻訳辞書を構築します(図 3)。この翻訳辞書は、たとえば、「私」と「I」が対訳関係にあることを記述しています。さらに、通常の辞書と異なる点として、どのくらいの確率でこれらが対応関係にあるかも記述しています。更に、単語だけではなく、フレーズの対訳関係も大量に格納しています。

この辞書を使って日本語を英語に翻訳する方法の概略は、まず、日本語文をフレーズに分割します。そして、そのフレーズを英語に翻訳して、最後に、翻訳したフレーズを並べ替えて英語の語順にします。ここで、もちろん、個々の日本語のフレーズには複数の英語のフレーズが対応しますし、日本語文をフレーズに分割する方法も多量にあります。したがって、自動翻訳が出力可能な英語文の数は無数にあります。この無数の翻訳候補の中から、前述の対訳関係の確率等を利用して、最適な英文を選択します。

もっとも、実際に行われている方法は、もっと複雑です。たとえば、英日自動翻訳の場合には、英語と日本語の構造差が大きいので、あらかじめ入力文の英語を構造解析して、日本語の語順に近くなるように英語の語順を変更してから、上述の方法で翻訳したりします。

自動翻訳の研究においては、これまで、ほぼ 10 年に 1 回の割合で、ブレークスルーとなる研究が起きていますので、ここ数年で、次のブレークスルーが起きるのではないかと思います。私たちは、そのブレークスルーを NICT から起こすように研究しています。