多くの言語対に対応する統計的機械翻訳システム

大量の対訳用例を用いて，任意の言語対の翻訳システムを自動構築

東芝は，20年以上にわたる研究開発により，日本語，英語，中国語間の相互翻訳を行う，高精度な機械翻訳システムを実現しました。更に精度を高め，より多くの言語対(注1)に対応するために，従来の規則ベース方式に代わる統計的機械翻訳(SMT：Statistical Machine Translation)の研究開発を，2003年から行っています。

大量の対訳用例から自動学習を行うSMTの精度は，用意した対訳用例の質と量に大きく依存します。当社は，目的にかなう対訳用例が十分に用意できないときにも高い翻訳品質を確保するために，ドメイン(対象分野)適応技術とピボット翻訳(第3の言語を介した2段階翻訳)技術を開発し，世界トップレベルの翻訳精度を実現しました。

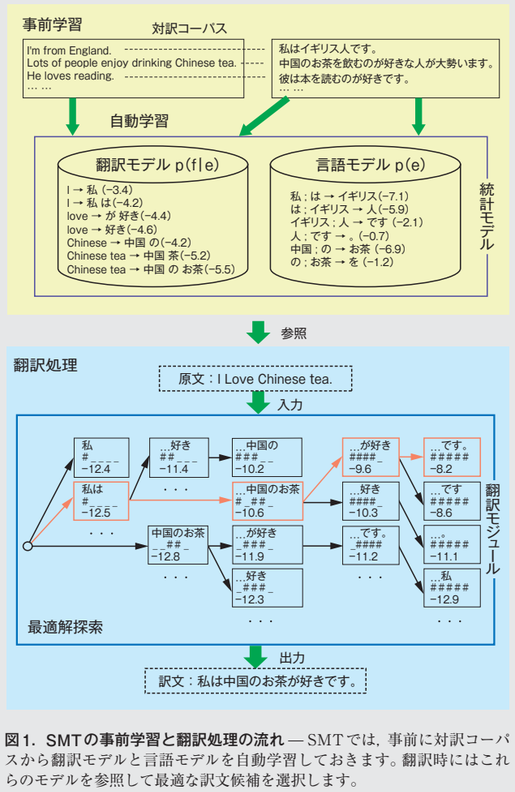

図1．SMTの事前学習と翻訳処理の流れ — SMTでは，事前に対訳コーパスから翻訳モデルと言語モデルを自動学習しておきます。翻訳時にはこれらのモデルを参照して最適な訳文候補を選択します。

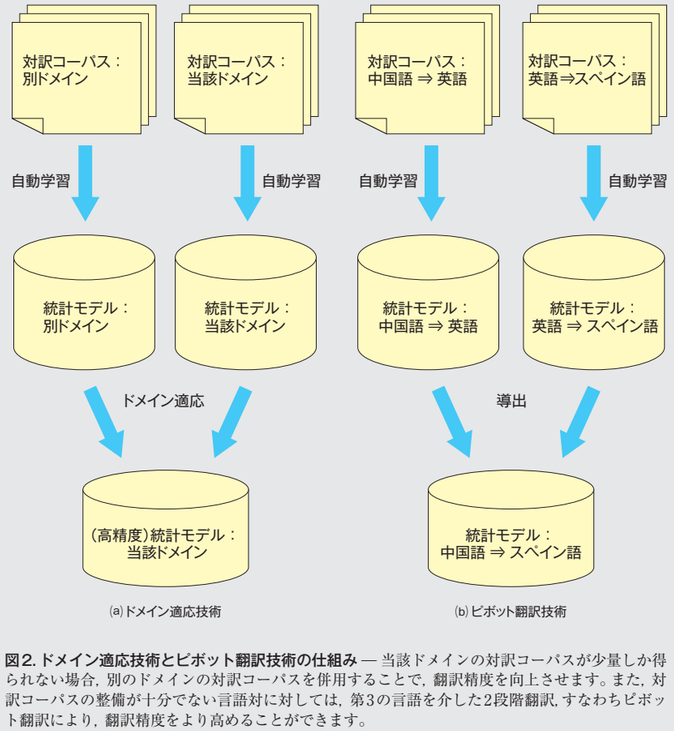

図2．ドメイン適応技術とピボット翻訳技術の仕組み — 当該ドメインの対訳コーパスが少量しか得られない場合，別のドメインの対訳コーパスを併用することで，翻訳精度を向上させます。また，対訳コーパスの整備が十分でない言語対に対しては，第3の言語を介した2段階翻訳，すなわちピボット翻訳により，翻訳精度をより高めることができます。

SMT開発の背景

インターネットの普及とグローバル化の進展により，機械翻訳はあらゆる層のユーザーにとって重要かつ身近な技術となりました。従来主流であった規則ベース機械翻訳(RBMT：Rule-Based Machine Translation)では，大規模な対訳辞書や翻訳規則の開発に時間とコストがかかるため，対象分野の拡大や，新たな言語対への対応が容易ではありません。

これに対して，大規模かつ高品質な対訳用例があれば自動的に高精度な翻訳システムを構築できることから，時間とコストのかかる辞書や規則の開発が不要なSMTが近年着目されています。

東芝は，日本語，英語，中国間の相互翻訳を行う高精度なRBMTシステムを既に実現しており，更に精度を高め，より多くの言語対に対応するため，SMTの研究開発に取り組んでいます。

(注1) 翻訳対象となる言語の組合せ。

SMTの仕組み

SMTの考え方は情報理論に由来します。SMTは，原文fが与えられると，条件付き確率p(e|f)を最大にする訳文eを出力します。実際には，p(e|f)の代わりに，これを変形したp(f|e)p(e)を最大化します。ここで，p(f|e)は翻訳モデル，p(e)は言語モデルと呼ばれ，確率値は大量に収集した言語データ(コーパス)から計算されます。

典型的なSMTシステムの事前学習と翻訳処理の流れを**図1**に示します。

事前学習では，対訳コーパスから翻訳モデルと言語モデルを自動学習します。通常言語モデルはn単語の連続出現確率で近似され，訳文側のデータから算出します。一方，翻訳モデルの学習には，原文と訳文の組が必要です。翻訳モデルは様々なものが提案されていますが，当社は，比較的成熟した“句に基づく”モデルをシステム開発で採用するとともに“構文に基づく”モデルの研究も行っています。

翻訳処理において，p(f|e)p(e)を最大化する訳文候補を選ぶ処理を，情報理論の用語に倣ってデコーディングと呼びます。訳文候補の数は膨大であるため，デコーダの実装にあたっては，翻訳品質と処理時間のトレードオフを考慮する必要があります。

実用化のための技術

SMTの性能はコーパスに依存します。しかし，目的にかなうコーパスを入手できないこともあるため，当社は，ドメイン適応技術とピボット翻訳技術を開発しました(**図2**)。

特定ドメイン向けのSMTシステムを開発する際に，当該ドメインの対訳コーパスが少量しか得られないことがあります。その場合，別のドメインの対訳コーパスを併用することで翻訳精度を高めることができます。また，対訳コーパスに比べて比較的入手が容易な単言語コーパスやドメイン対訳辞書を利用して，ドメイン適応を行うこともできます。

対訳コーパスの整備が十分でない言語対も数多くあります。その場合，第3の言語を介した2段階翻訳，すなわちピボット翻訳が有効なことがあります。例えば，中国語ースペイン語間の大規模な対訳コーパスはありませんが，中国語ー英語，英語ースペイン語間は豊富にあり，英語を介したピボット翻訳のほうが直接翻訳より高い精度が得られます。

世界トップレベルの翻訳精度

当社は，優れたSMTツールキットと高品質な言語資源(対訳コーパス，単言語コーパス，及び各種辞書)を整備し，これらを用いて構築したSMTシステムにより，世界トップレベルの翻訳精度を実現しました。IWSLT(International Workshop on Spoken Language Translation)という評価型ワークショップの五つのタスクに参加し，三つのタスクで単独1位，残り二つのタスクで1位グループという実績があります。

今後の展望

今後，異なる翻訳方式の組合せによる翻訳精度の向上も期待できます。当社が開発したRBMTシステムは，言語対は限られるものの，大量に蓄積された貴重な言語知識により高い翻訳精度を実現しています。更に，用例ベース機械翻訳(EBMT：Example-Based Machine Translation)もSMTに匹敵する性能を持っています。

これらとSMTの組合せによる翻訳精度の向上を予備実験で既に確認しており，今後，優れた“ハイブリッド機械翻訳”システムを目指して更に研究開発を進めるとともに，SMTの特長である多言語展開も同時に進めていきます。

王　海峰
東芝中国社
研究開発センター
副所長

出羽　達也
研究開発センター
知識メディアラボラトリー主任研究員