目覚ましい発展の中の小さな発見

Graham Neubig　奈良先端科学技術大学院大学 情報科学研究科

〔受賞論文〕
Joint Phrase Alignment and Extraction for Statistical Machine Translation
Graham Neubig（京都大学 情報学研究科），Taro Watanabe, Eiichiro Sumita（情報通信研究機構），Shinsuke Mori, Tatsuya Kawahara（京都大学 情報学研究科）
Journal of Information Processing, Vol.20, No.2, pp.512-523（2012）

このたび，本会論文賞をいただくことになり，大変光栄に思っております．

本論文は日本語や英語のような自然言語の間で自動的に翻訳する機械翻訳について述べたものである．機械翻訳は長年の夢であり，近年の目覚ましい発展でようやく現実味を増してきている．この発展を支えているのは大量の研究開発結果であり，毎年数百本の論文が発表される．最先端のシステムではこの研究結果を組み合わせ，精度の向上を実現しているが，システムの複雑化も目立つ．

システムの複雑化の象徴とも言えるのは，対訳データから翻訳に利用される翻訳ルールを学習する過程である．まず対訳データから，対訳文の中でどの単語がどの単語に翻訳されるかを発見する「単語アライメント」と，単語ごとの対応から翻訳ルールを構築する「ルール抽出」がある．従来，この2つの段階はヒューリスティックスと確率モデルを組み合わせ，そのからみ合いでモデルの透明性が損なわれていた．本研究では2段階化を除外し，1つの確率モデルで対訳データからルールを直接構築する手法を開発することで，この複雑さの解消を目指した．

すっきりした確率モデルを実現するべく試行錯誤を繰り返し，ようやく鍵が意外なところに見つかった．この鍵は毎年大量に発表される機械翻訳の論文ではなく，15年前の一見関係ない博士論文にあった．論文では日本語文のような，明示的な単語境界を利用しない文字列を自動的に単語に分割することをテーマにしていた．肝になっている手法は，言語の再帰的構造を二分木上の確率モデルとして表すことで，**図-1 a)** のように文字や単語の隣同士を組み合わせてより長い意味的単位を発見する．

この論文を発掘してから研究にすぐに取り掛かった．言語の再帰的構造を表す木は機械翻訳と相性が良く，**図-1 b)** のように，2言語間を渡る木は単語アライメントの研究でも広く利用されていた．このアライメントの従来法と発掘した論文のアイディアを融合し，**図-1 c)** のように2言語間の単語をより長いフレーズに組み合わせるモデルを提案した．さらに，15年前のモデルを近年提案されている統計モデルや解析アルゴリズムで磨き，複雑さを解消しながら高い翻訳精度を実現する学習アルゴリズムに仕上げた．

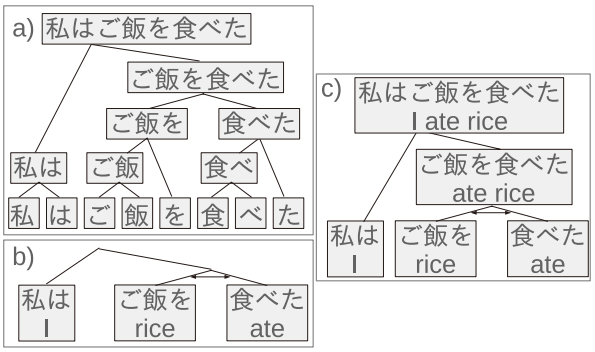

図-1　a）再帰的な単語分割モデル　b）単語アライメントモデル　c）両者を組み合わせた提案モデル

研究開発が猛烈なペースで進む情報学分野の中で，自分の研究に関連深い論文を追っていくことに精一杯になりがちである．しかし，本研究の経験から，時々一歩引いて広い視野で周辺の研究を眺めることの重要さが身にしみて分かった．この精神を忘れずにこれからも研究や勉強に励みたい．

（2013年5月10日受付）

Graham Neubig（正会員）　neubig@is.naist.jp
2005年米国イリノイ大学工学部コンピュータ・サイエンス専攻卒業．2010年京都大学大学院情報学研究科修士課程修了．2012年同大学院博士後期課程修了．同年奈良先端科学技術大学院大学助教．自然言語処理に関する研究に従事．