早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論　文　題　目

A Study of Efficient Bidirectional Latent Variable Models in Machine Translation

申　請　者
Hao WANG

情報生産システム工学専攻
用例翻訳•言語処理研究

2019 年 02 月

統計的機械翻訳（SMT）とニューラル機械翻訳（NMT）の機械翻訳パラダイムでは、翻訳は機械学習問題として扱われているため、条件付き確率モデルが構築される。源言語から目的言語へのモデルとなり、一方向モデルである。従って、言語間の相違は完全には反映されない。いずれのパラダイムであっても、英語と日本語のような離れた言語対の翻訳は困難である。源言語と目的言語間の相違は、単語への分割結果（分かち書き）、単語間の対訳関係（アライメント）、単語の順序（並べ替え）に位置付けられる。従来技術では単言語モデルや一方向モデルを使用するため、語彙間、単語アライメント間、構文構造間の相当する共通点を部分的に見落としており、モデルサイズが大きくなり学習所要時間も長くなる。

本論文では、双方向潜在変数モデルを使用することで、より効率良く両言語での相当する部分を捉えることを目標とする。3 つの異なる機械翻訳アプローチにおいて効率性の高い双方向潜在変数モデルを提案し構築する。ここでの効率性とは（a）最先端技術と同等またはそれ以上の翻訳品質となることを前提とし（b）構築されるモデルのサイズの縮小、（c）学習所要時間の削減、またはいずれかとする。英日の離れた言語対に適用することにより、提案されるモデルの効率性の根拠は本質的に双方向潜在変数モデルにあることを明らかにする。

以下に、本学位論文の構成と各章ごとの評価を述べる。

第 1 章「Introduction」と第 2 章「Background」では、機械翻訳パラダイムについて述べ、潜在変数モデルと双方向モデルの基礎概念を導入する。SMT 及び NMT に関する必要な概念について述べている。

第 3 章「Exploiting bidirectional latent variables in phrase-based SMT: phrasal alignment」では、句に基づく統計的機械翻訳における重要なパラメータである句アライメントの双方向生成手法を提案する。最先端技術では、まず単語アライメントを計算し、EM アルゴリズムを使用する（複雑さの低い IBM モデル[Brown et al. 1990]）。次に、単語アライメントに基づき、句アライメントを推定するため、より複雑さの高い IBM モデル[Brown et al. 1993]の計算を行う。最後に、上記の学習は一方向手法であるため、各方向アライメントに基づきヒューリスティック的な手法で双方向翻訳テーブルを生成する[Och and Ney 2003]。ここで訓練の所要時間が長いと生成される翻訳テーブルが大きくなる問題がある。[Lardilleux et al. 2012]では、単語アライメントの推定値に基づく階層的双方向句アライメント（HSSA）手法を提案し、生成時間を削減させることができた。しかし、パラメータの初期値に鋭敏性があり、最先端技術の翻訳品質を超えていない。

本論文では、HSSA に基づくより頑丈な句アライメント手法を提案する。まず、単語アライメントをより安定的に推定するため、EM アルゴリズムに変分ベイズ法[Riley and Gildea 2012]を導入する。次に、HSSA の句アライメントの信頼性が高いため、複雑さの低い IBM モデルにその信頼性に影響の与えない近似を導入す

る。最後にHSSA手法にビームサーチを導入することにより、双方向バランスをより安定させる。したがって、出力句アライメントの品質と的確さを向上させた。また、複雑さの高いIBMモデルの計算は不要となり、結果として加速させた。

全てのIBMモデルとヒューリスティック的な翻訳テーブル生成から成る最先端技術と比較すると、提案した手法は複数の言語対での実験で、(a) 翻訳品質を維持しながら (b)生成された翻訳テーブルのサイズを縮小させ（平均的に５０％)、(c)学習所要時間を大幅に削減させた（最先端技術の平均時間の４％)。速度の速い fast_align [Dyer et al. 2013]というアライナーと比較すると、英日の実験で(a)統計的に有意に翻訳品質を向上させ（+1.37BLEU、p 値<0.01)、(b) 半分のサイズの翻訳テーブルを生成しながら(c)学習所要時間は同等であった。

第４章「Exploiting bidirectional latent variables in syntax-based SMT: syntactic representation」では、2 つの言語の構文構造の対訳関係を表す双方向潜在変数モデルについて述べている。構文解析器のない言語にはもちろん、構文構造が顕在な変数モデルの直接的な構築は不可能である。階層統計的機械翻訳[Chiang 2007]のような構文に基づくSMT手法では、構文解析器が不要であるが、かわりにより複雑な翻訳モデルの計算が必要となる。ただ、長距離に渡る並べ替え問題が生じてくる。その問題を把握するため、bracketing transduction 文法（BTG) [Wu 1997] というエレガントな双方向形式文法が提案されてきた。機械翻訳では、BTGの使い方が二つある：一つは学習する前（事前並べ替え)、もう一つは翻訳と同時に（デコーディングの際）であるが、デコーディングは従来実装困難とされてきた。[Nakagawa 2015]の事前並べ替え手法は最先端の結果に至った。しかし、問題として、単語アライメントと構文アライメントに起きうるエラーに対し重大な鋭敏性があることが知られてある。

本論文では、エラーにさほど鋭敏性のない双方向モデル構築手法を提案し、それに基づく、双方向デコーディング手法を提案した。複数の機械学習技術（ミニバッチ、k-best list 等）を用いることにより、訓練の際、並べ替えのより的確な構文を選択すると同時に安定性を改善し、両言語での共通構文の量を増大させた。並べ替えスコアで評価すると、最もよい場合には、fuzzy reordering score (FRS) では+2.00、normalized Kendall’s tau (NKT)では+0.42 の統計的に有意な改善が見られ、単語と構文のアンバランスを減少させた。

事前並べ替えの英日両方向翻訳実験において、(a) 統計的に有意な改善が見られ（それぞれ+0.5、+0.8BLEU、p 値<0.01)、(b) モデルのサイズは変わらず (c) 訓練所要時間を削減させた（４倍の高速化)。また、デコーディング実験では、英日翻訳においては、最先端技術のMOSESデコーダと比較すると、(a) 統計的に有意な改善が見られた(+0.43 BLEU、p 値<0.05)。それに加え、(b) 並べ替えモデルのサイズを縮小させ(-20%まで)、デコーディング時間には影響は及ばなかった。

第５章「Exploiting bidirectional latent variables in Seq2seq NMT: sub-word

segmentation」では、分かち書きで得られた語彙の双方向潜在変数モデルについて述べている。分かち書きはアジア言語（中国語、日本語）の語彙を自動的に決定するために使用される。従来の教師あり分かち書き器では、出現頻度の低い単語が多く生成され、Seq2seq NMT ではその単語の翻訳は不可能である。出現頻度の低い単語の量も異なる分かち書き器で差異がある。

本論文では、出現頻度の低い単語問題と同時にその量の差異の問題を解くため、最小記述長（MDL）の原則に基づく教師なし双方向サブワード分割法を提案する。通常の MDL に二つの制約を課す：一つは、有限な語彙、もう一つは最小出現頻度である。その結果、文字数の多い言語において、有限なサブワード集合を生成し、それぞれのサブワードは最低出現頻度を超え、あくまでも両言語に共通するサブワードを出力できた。共通するサブワードが存在するため、Seq2seq NMT モデルのエンコーダーとデコーダーの単語埋め込み層は統一させることができた。

日中の NMT 実験において、従来技術の Juman [Matsumoto et al. 1994]および Stanford Segmenter [Chang et al. 2008]と比較すると、提案した手法を使用することによって、(a)翻訳品質では統計的に有意な改善が見られ(+1.7、+2.4 BLEU、p 値<0.01)、(b)語彙サイズは縮小した(10 倍小さい)。(c)Juman および Stanford Segmenter は既製モデルとして使用するため時間比較は不可能である。SentencePiece モデル[Wu et al. 2016]やバイト対符号化[Sennrich et al. 2016])の 2 つのサブワードモデルと比較すると、日英と日中の両方向翻訳実験において、（a）翻訳品質に変わりはなかった。（b）語彙量が同じである条件の下では、（c）訓練所要時間が 3～10 倍速くなった。

第 6 章「Conclusion and contribution」では、上記の機械翻訳分野への貢献を総括し、本論文の結論及び今後の展開方向について述べている。

以上を要約すると、本論文では、機械翻訳で言語間相違問題がある三つの過程(分かち書き、アライメント、並べ替え)に、従来の潜在変数モデルにある初期値の鋭敏性の問題と、一方向モデルでは的確的に捉えられていない言語間相違の問題に対して、より頑丈な双方向潜在変数モデルを提案し構築した。双方向モデルであるからこそ、より的確に両言語の共通部分を捉えることができた。その結果、従来技術の翻訳品質を維持、またはそれを上回り、場合によっては大幅にモデルのサイズが縮小され訓練所要時間を削減できた。特に英日の離れた言語対で翻訳品質の向上ができた。これらの成果は機械翻訳分野の発展に寄与するところ大である。よって、本論文は博士（工学）の学位論文として価値あるものと認める。

2019 年 2 月 5 日

審査員

主査　早稲田大学　教授　博士 (工学) (グルノーブル大学) ルパージュ・イヴ

早稲田大学　教授　博士 (情報工学) (九州工業大学) 古月　敬之

早稲田大学　教授　博士 (工学) (九州大学) 岩井原　瑞穂