後藤　功雄　（黒橋教授）

「Word Reordering for Statistical Machine Translation via Modeling Structural Differences between Languages」

（統計的機械翻訳のための言語構造の違いのモデル化による語順推定）

平成 26 年 5 月 23 日授与

機械翻訳はグローバル化が進む現代においてニーズが増加している。近年中心的に研究されている機械翻訳は、大規模な対訳コーパスから翻訳知識を自動獲得して利用する統計的機械翻訳（SMT）である。機械翻訳は主に「訳語選択」と「語順推定」という 2 種類の処理が必要である。語順が大きく異なる言語間の翻訳では長距離の語順並べ替えが必要であるが、従来の SMT 手法で利用しているフレーズレベルの局所的な情報では長距離の語順並べ替えを高精度に行うことは困難であった。図 1 に語順が大きく異なる言語である日本語と英語の対訳文の例を示す。これらの言語間では、長距離の並べ替えを含む複雑な並べ替えが必要であることが分かる。

入力文から目的言語の構造（syntax）に従った語順を推定するには、原言語と目的言語の構造の違いに対するモデルが有用である。世界には多くの言語があるが、ほとんどの言語では高性能な構文解析器がない。また、目的言語で構文解析器が使える翻訳のニーズが大きい。そこで、本論文は、構文解析器を必要としない場合および目的言語の構文解析器を利用する場合において、言語構造の違いを既存手法より適切にモデル化することで、SMT における語順推定を改善する手法を提案した。構文解析器を必要としない SMT の語順推定モデルのモデル化手法として、単語列ラベリングに基づく手法を提案した。このモデルは、入力文の翻訳過程において、最後に翻訳した位置と次に翻訳するべき位置候補の範囲をラベル系列で表現することで、動詞句や名詞句などの特徴を捉える。これによって、言語間の構造の違いを近似的にモデル化した。目的言語の構文解析器を利用する SMT の語順推定手法として、2つの手法を提案した。1つは構文解析を利用するポストオーダリング手法である。日英の構文構造の違いは大きく、言語間で同期がとれない部分が多いために、モデル化が難しい。それに対して、この提案手法では、目的言語の構造を原言語風の構造に変換した構造と変換前の構造との違いをモデル化することで、言語構造の違いのモデル化を容易にした。もう1つは目的言語の構文解析器を用いたプレオーダリング手法である。射影によって目的言語の構造となるべく同期がとれる原言語の構造を構築することにより、言語構造の違いのモデル化という課題を扱いやすくした。

語順の推定には言語構造が重要である。言語構造の同定の性能は改善の余地があるため、構造の同定の改善および構造の解析誤りに頑健な翻訳手法の開発が今後の課題である。

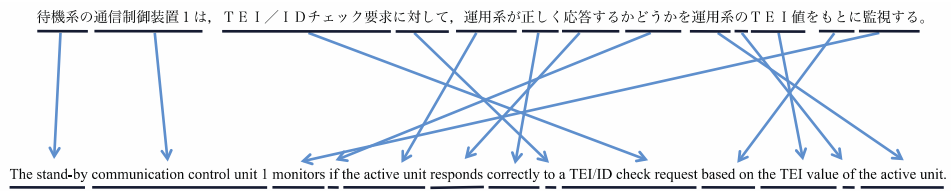

図 1　語順が大きく異なる言語間（日英）の語順の違いの例