統計的機械翻訳における語義に基づく大局的な情報を用いた語彙モデルに関する研究

A Study on Global Lexicon Model Based on Word Sense in Statistical Machine Translation

学籍番号：201221581
氏名：大山 鉄郎
Tetsuro OHYAMA

統計的機械翻訳では，入力文に対する複数のモデルの評価値から，出力文が決定される．モデルの一つとして，原言語の広い範囲の文脈から目的言語を推定する，大局的な情報を用いた語彙モデルがある．このモデルは，ある単語の目的言語の推定に，その単語が出現した文に含まれる他の単語の表層形を素性として用いる．広い範囲の文脈を考慮した翻訳ができるが，素性の組み合わせは膨大なものになり，適切な評価値が与えられないことも多い．

本研究では，語義に基づく大局的な情報を用いた語彙モデルを提案する．提案手法は，非活用語である名詞と副詞を語義に置き換えることで，同じ意味を持つ単語を，同じ素性と見なす．これにより，語義が同じ単語は，同じ文脈に出現するものとして扱われる．語義を用いることで，単語の表層形のみを使用する場合よりも，柔軟に文脈を考慮することができる．

提案手法は，原言語と目的言語から生成する語義と，原言語の表層形とを用いて目的言語の単語を推定する．単語から語義の変換には WordNet を用いる．WordNet は，英語と日本語の単語が語義でまとめられている．単語は複数の語義を持つため，文脈において不適切な語義に変換されることがある．提案手法では，原言語と目的言語それぞれで単語から語義への変換を行い，その積集合を取ることで，不適切な語義を素性に含まないようにしている．

提案する語彙モデルは，原言語と目的言語から生成した語義と表層形の集合 $\boldsymbol{s}$ が与えられた際の目的言語の文 $\boldsymbol{e}$ の確率を，単語 e の条件付き確率 $p(e|\boldsymbol{s})$ の積で表現する．$p(e|\boldsymbol{s})$ は，最大エントロピーモデルであり，e に対して $\boldsymbol{s}$ の要素それぞれに割り振られた確率の和を，全ての単語に対して与えられた $\boldsymbol{s}$ の要素に割り振られた確率の和で割った値で表す．

提案手法は，旅行会話文と特許文から構築された 2 種類のコーパスを用いた，統計的機械翻訳システム Moses による日英翻訳において，ベースラインより，評価尺度 $BLEU$ で 1.30 と 0.09 ポイント，$RIBES$ で 0.34 と 0.37 ポイントの向上を示した．語義に基づく大局的な情報を用いた語彙モデルは，文脈を考慮すると不適切である訳語を改善することを確認した．また，単語やフレーズの翻訳において，文全体の整合性を保つように訳語を選択することを示した．本研究では，全ての単語を語義に変換したが，推定に語義が不要な単語もある．それらを判別することによって，提案手法の更なる改善が見込める．

研究指導教員：佐藤 哲司
副研究指導教員：関 洋平