The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ニューラル機械翻訳における単語予測の重要性について

竹林　佑斗 *1　Chenhui Chu*1　荒瀬 由紀 *1　永田　昌明 *2

*1大阪大学　*2日本電信電話株式会社

1. 序論

本稿では、人手または統計的手法により自動作成された対訳辞書を利用するニューラル機械翻訳 (NMT) のデコーディング法を提案する。提案手法は、入力文から出力文に出現すべき単語集合を予測する方法と、予測された単語集合が生成されることに報酬を与えるようなデコーディング時のスコアの計算法から構成される。本稿では、科学技術論文の日英翻訳タスク (ASPEC-JE) を対象として、予測単語集合の再現率と適合率を様々に変化させるシュミレーションにより提案手法の有効性を検証した。例えば、予測単語集合としてオラクル単語集合 (正解訳の単語集合) を与えた場合、提案したスコア計算法を用いることにより BLEU スコアが 7.74 ポイント向上することを確認した。

2. 単語集合予測に基づく単語レベル報酬モデル

ここで、本稿の提案法を述べる。提案法は、大きく二つの構成要素からなる。

1. 入力系列から出力系列に出現する各単語の集合（目的単語集合）を予測するモデル
2. 目的単語集合を用いてデコーダが選択する単語を制御する機構

図 1 にデコーディングステップ j における提案法の振る舞いを示す。まず入力系列から目的単語集合を予測する。予測された単語の生成確率に報酬を与えることで、その単語を生成されやすくする。各ステップにおいて、以下の式を用いて単語の生成確率を増加させる。この式は [Li 17] の，出力系列の長さを制御するモデルを参考にした。

$$y_j = \arg\max_y (\log P(y_j|y_{<j}, \mathbf{x}) + \lambda d_y), \qquad d_y \in \mathbf{D}_{f2e}, \quad (1)$$

ここで $\mathbf{D}_{f2e}$ はは対訳辞書とし、λ は報酬の重みである。d_y は、 $y \in D_{f2e}$ ならば 1、そうでなければ 0 であるような関数である。

3. 単語集合予測のシミュレーション

本稿では、提案手法の 2 番目の構成要素の有用性を検証するため実験を行なった。まず参照文に対する単語予測の適合率 (precision) と再現率 (recall) が共に 100%であるオラクル単語集合を作成し、提案手法の効果の上限値を調査した。さらに、オラクル単語集合への単語の追加および削除の処理を加えることで、適合率と再現率をそれぞれ変化させて、予測結果のシミュレーションを行いその時の翻訳性能を測った。その結果、どのような適合率と再現率を持つ目的単語集合を予測できれば、提案手法を用いることでベースラインを大きく上回る結果が得られるかが明らかとなった。

3.1 対訳データと翻訳ツール

ASPEC-JE を用いて提案手法の有効性をシミュレーションする。このデータでは 300 万文のうち上位 200 万文を訓練データとし、 開発データとして 1,790 文、テストデータとして 1,812 文を用いた。提案モデルのベースラインとして「深層学習による自然言語処理」*1 の付録として公開されている mlpnlp-nmt*2 (注意機構付き LSTM エンコーダデコーダモデル) を用いた。このモデルは WAT-2017 の ASPEC-JE/EJ タスクにおいて人手評価でもっとも優れた性能を納めたものである [Nakazawa 17]。提案手法はこのモデルのデコーダ部を上述の式の通り書き換えることで実装した。ハイパーパラメータの設定は [Morishita 17] に従った。

3.2 オラクル単語集合からの精度調整

オラクル単語集合は参照文の単語を目的単語集合とすることで得た。シミュレーション用の擬似予測単語集合を作成する際はオラクル単語集合に下記の処理を加えることで得た。再現率を減少させる際は、オラクル単語集合から複数の単語を無作為に削除し True positive を減少、および False negative を増加させることで調整した。適合率を減少させる際は、オラクル単語集合に複数の単語を追加し False positive を増加させることで調整した。近似予測をできる限り現実的にするため、追加する単語は GIZA++*3 でワードアライメントを取った結果、入力系列の単語から出力系列単語への単語翻訳確率の高い単語から追加した。

3.3 実験結果

オラクルの単語集合を用いた時の BLEU スコアは、35.64 であった。これはベースラインシステムの BLEU スコアの 27.90 から+7.74 ポイントの向上であり、ベースラインの翻訳性能を有意に上回っている ($p < 0.01$)。図 2 に擬似予測単語集合を用いたシミュレーションの結果を示す。各点は擬似単語集合の適合率と再現率をそれぞれ 10% – 90% まで 10% 刻みで変化させた時の、BLEU スコアを表している。*border* は、その線以上の結果はベースラインを有意に上回っているという境界線を示している。したがって、*border* 以上の適合率と再現率をもつ単語集合を入力系列から予測できれば、我々の提案手法を用いることでベースラインを優位に上回る性能が得られる。結果より、適合率よりも再現率のほうが翻訳性能の向上に大きく貢献しており、例えば適合率が 10%であっても再現率が 50%あれば翻訳性能が大きく向上するが再現率が 10%であると適合

連絡先: 竹林　佑斗、大阪大学、大阪府吹田市山田丘 1-5、takebayashi.yuto@ist.osaka-u.ac.jp

∗1 http://www.kspub.co.jp/book/detail/1529243.html
∗2 https://github.com/mlpnlp/mlpnlp-nmt
∗3 http://code.google.com/p/giza-pp

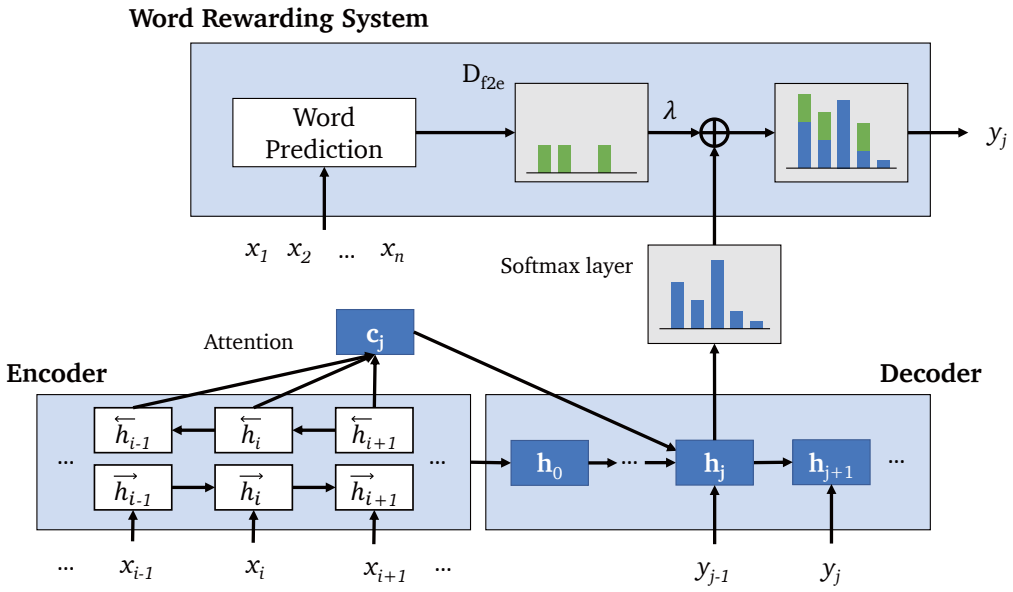

図 1: デコーディングステップ j における単語レベル報酬モデル。各デコーディングステップにおいて、予測された目的単語集合内の単語生成確率に報酬を与えることで、生成されやすくする。

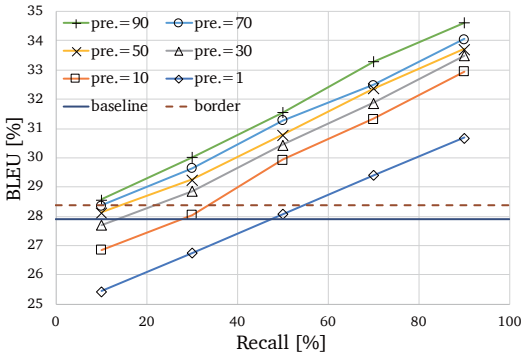

図 2: 擬似予測単語集合を用いて適合率と再現率を変化させた時の BLEU スコア。*boder* は翻訳結果がベースラインと差があるものの境界を示しており、*border* より上の結果はベースラインを優位に大幅に上回っている ($p < 0.01$) ものである。

率が 70%より多くなければ *border* を超えないという結果が得られた。

4. 結論

本稿では、出力系列を制御する手法として、単語レベル報酬モデルを提案した。このモデルは辞書などの外部知識を NMT モデルに適用するものであり、NMT モデルを再訓練する必要がないのが特徴である。実験では擬似予測を行い適合率と再現率を変化させて目的単語集合をシミュレーションし、それを用いた時の BLEU スコアを求めて、ベースラインと統計的有意差があるかどうかを検定した。その結果、予測モデルで作成する目的単語集合に求められる適合率と再現率が明確になった。今後の課題としては、実際に目的単語集合を予測して本稿のシミュレーション結果を満たす結果が得られるかを検証する。

参考文献

[Li 17] Li, J., Monroe, W., and Jurafsky, D.: Learning to Decode for Future Success, *CoRR*, Vol. abs/1701.06549, (2017)

[Morishita 17] Morishita, M., Suzuki, J., and Nagata, M.: NTT Neural Machine Translation Systems at WAT 2017, in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pp. 89–94, Taipei, Taiwan (2017), Asian Federation of Natural Language Processing

[Nakazawa 17] Nakazawa, T., Higashiyama, S., Ding, C., Mino, H., Goto, I., Kazawa, H., Oda, Y., Neubig, G., and Kurohashi, S.: Overview of the 4th Workshop on Asian Translation, in *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pp. 1–54, Taipei, Taiwan (2017), Asian Federation of Natural Language Processing