# 単言語コーパスと逆翻訳を用いたエンコーダー・デコーダーの訓練法

今村 賢治　　藤田篤　　隅田 英一郎
国立研究開発法人 情報通信研究機構
{kenji.imamura,atsushi.fujita,eiichiro.sumita}@nict.go.jp

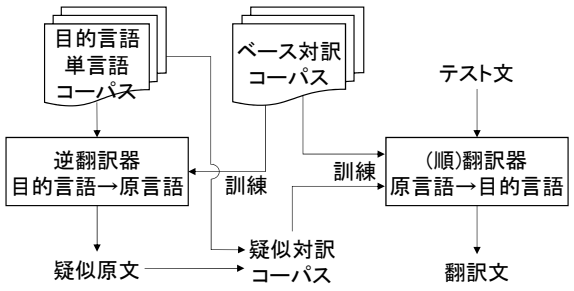



図 1: 本稿の方式のフロー

## 1　はじめに

近年の機械翻訳は，エンコーダー・デコーダー方式のニューラル機械翻訳 (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) が主流となってきている．これは，入力文（原文）をエンコーダーによって，状態と呼ばれる数値ベクトルに符号化し，デコーダーが状態に基づいて翻訳文を生成する方式である．エンコーダー・デコーダー方式は従来の統計翻訳方式に比べ，高品質な翻訳文を生成することが可能であるが，この訓練には大量の対訳文が必要である．しかし，大規模対訳コーパスは，単言語コーパスに比べ，一般的に入手が難しい．

この問題に対し，Sennrich et al. (2016a) は，目的言語の単言語コーパスを原言語へ逆翻訳して疑似対訳文を生成し，対訳コーパスと混合して訓練する方法を提案した（図 1）．この方法の利点は，疑似対訳文といっても，目的言語側は人が作成した正しい文が使用されるため，デコーダーは正しく訓練されることである．そのため，単言語コーパスから言語モデルを構築する方法に比べても，安定した精度向上が可能である．しかし，エラーを含む疑似原文で訓練するため，エンコーダーの精度向上にはあまり寄与していない可能性がある．

そこで本稿では，Sennrich et al. (2016a) の方法を拡張し，目的言語の単言語コーパスでエンコーダー（アテンションを含む）を強化して，翻訳器全体の精度向上を試みる．本稿の提案方式は，逆翻訳によって疑似原文を生成する際，サンプリングによって複数の文を生成して訓練に用いる．複数の疑似原文を用いることにより，以下の効果を期待する．

- 個々の疑似対訳文に含まれるエラーを平均化して，影響を軽減させる．
- 機械翻訳で作成した対訳は，どうしても単調（いつも同じような訳）になるため，複数の疑似原文で，人間に近い多様性を確保する．

実験では，1 文あたりの疑似原文数を増やすにしたがい，翻訳品質は向上した．人手で逆翻訳をした場合（つまり，通常の対訳文を追加した場合）と比較しても，近い品質が出せることがわかった．

## 2　提案方式: 単言語コーパス逆翻訳によるエンコーダーの訓練

### 2.1　疑似原文生成

提案方式で使用する逆翻訳器は，比較的小規模な対訳コーパス（ベース対訳コーパスと呼ぶ）で訓練された NMT である．目的言語の単言語コーパスから 1 文ずつ取り出し，逆翻訳器で原言語文を生成する．ただし，逆翻訳器は尤度の高い文を出力するのではなく，ランダムサンプリングで生成する．つまり，デコーダーは 1 単語を出力する際，出力単語の事後確率分布を生成するが，そこから取り出す単語を，確率最大のものではなく，確率分布で重み付けされたサンプリングで決定する．

$$y \sim Pr(\boldsymbol{y}_n|\boldsymbol{y}_{<n}, \boldsymbol{x}) \tag{1}$$

ただし，$y$ は出力単語，$\boldsymbol{y}_n$ はその位置での単語分布，$\boldsymbol{y}_{<n}$ は出力単語履歴，$\boldsymbol{x}$ は入力単語列である．

表 1 は，疑似原文の生成例である．逆翻訳器が出力する対数尤度でソートしてある．これを見るとわかるように，人手逆翻訳（参照訳）と同一，または近い疑似原文が多く生成されている．一方，5 番目の疑似原

表 1: 疑似原文生成例（英日翻訳）

| 対数尤度 | 疑似原文 |
|---|---|
| -2.25 | what should i do when i get injured or sick in japan ? |
| -2.38 | what should i do if i get injured or sick in japan ? |
| -5.20 | what should i do if i get injured or illness in japan ? |
| -5.52 | what should we do when we get injured or sick in japan ? |
| -13.87 | if i get injured or a sickness in japan , what shall i do ? |
| 目的言語 | 日本で怪我や病気をしたときはどうすればいいのでしょうか？ |
| 人手逆翻訳 | what should i do when i get injured or sick in japan ? |

文は節が倒置されており，語順の観点からは，人手逆翻訳とはかなり異なる．この疑似原文は，尤度も低いため，N ベスト翻訳では通常出力されない文であるが，サンプリングであれば，このような多様性のある疑似原文も生成できる．

## 2.2 エンコーダーの訓練

生成された疑似原文は，目的言語文と対で疑似対訳文にして，(順）翻訳器の訓練に使用する．基本的にはベース対訳コーパスと疑似対訳コーパスを混合して学習を行う．

しかし，目的言語 1 文に対して複数の疑似原文を使っている場合，両者を単純に混合すると，疑似対訳が過度に学習される．この問題を避けるため，ベース対訳と疑似対訳の学習率を変えて学習する．具体的には，まず，ベース対訳と疑似対訳で個別にミニバッチ集合を構成する．そして，ベース対訳の学習率 $\eta$ に対し，疑似対訳の学習率を $\eta/N$ にして，ミニバッチに付与する．ただし，$N$ は，目的言語 1 文あたりの疑似原文数である．そして両者をシャッフルして学習する．

なお，単言語コーパスとベース対訳コーパスのドメインが同じ場合は，上記のとおり訓練すればよい．もし，両者のドメインが異なる場合は，この後にベース対訳で追加訓練して，ドメイン適応させることが望ましい．[1]

## 2.3 疑似対訳のフィルタリング

本稿の狙いの一つは，個々の疑似原文に含まれるエラーを軽減することである．エラーを直接軽減するには，疑似対訳の品質でフィルタリングする方法も考えられる．今回，複数の疑似原文（疑似対訳プールと呼ぶ）の中から，以下の 3 種類のフィルタリング方法を検討する．

- 疑似対訳プールから，尤度により選択する．ただし，尤度は疑似原文の長さで補正した値（本稿では長さ補正尤度 $ll_{\text{len}}$ と呼ぶ）を使用する (Oda et al., 2017).

$$ll_{\text{len}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_t \log Pr(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) + WP \cdot T \quad (2)$$

  ただし，右辺第 1 項は逆翻訳器が出力した対数尤度，$WP$ は単語ペナルティ（$WP \geq 0$），$T$ は疑似原文の単語数である．単語ペナルティは，開発セットにおいて，翻訳文と参照訳の長さがほぼ同じになるように設定する．
- 疑似対訳プールから，信頼度に基づいて選択する．信頼度は，機械翻訳文の品質推定タスクで用いられている指標 (Fujita and Sumita, 2017) を使用する．これは，機械翻訳文が許容可能かどうかを，SVM で学習し，文単位の信頼度として出力するものである．
- 疑似対訳プールから，ランダム選択する．疑似原文の生成数を減らした場合と同じ．

表 2: コーパスサイズ

| 種別 | | 文数 |
|---|---|---|
| 対訳 | ベース対訳 | 400,000 文 |
| | 開発 | 2,000 文 |
| | テスト | 2,000 文 |
| 単言語 (日本語) | GCP コーパス | 1,552,475 文 |
| | BCCWJ | 4,791,336 文 |

# 3 実験

## 3.1 実験設定

**コーパス**　本稿で用いたコーパスのサイズを表 2 に示す．対訳コーパスは，我々が開発した GCP コーパス (今村，隅田, 2018) を使用した．これは 10 言語のパラレルコーパスであるが，今回は日本語，英語，中国語対訳部分を用い，英日翻訳，および中日翻訳で実験した．GCP コーパスのうち，40 万文をベース対訳とし，残り約 155 万文を単言語コーパスとして用いた．同一コーパスをベース対訳と追加の単言語コーパスに分割した理由は，既存対訳を精度向上の上限として比較したいためである．

また，ベース対訳とはドメインが異なる単言語コーパスとして，現代日本語書き言葉均衡コーパス (BCCWJ)1.1 の，1024 文字以下の約 480 万文を使用した．

[1] 本稿では，追加訓練は行わなかった．

すべてのコーパスは，内部開発の形態素解析器で単語分割した．さらに，日本語，英語，中国語それぞれ独立に，ベース対訳から獲得したバイトペア符号化 (Sennrich et al., 2016b) ルールで，1.6 万のサブワードに分割して使用した．

**翻訳システム**　今回使用した翻訳器は，OpenNMT (Klein et al., 2017) である．これを，2.1，2.2 節対応に改造して使用した．

エンコーダーは 2 レイヤー Bi-LSTM (500+500 次元), デコーダーは 2 レイヤー LSTM(1,000 次元) とし，最適化は確率的勾配降下法 (SGD) を使用した．ベース対訳コーパスの学習率は，1.0 で 14 エポック，その後半減させながら 6 エポック学習した．ミニバッチサイズは 64 とした．

翻訳はビーム幅 10 で 10 ベスト翻訳を生成後，長さに基づくリランキング (Morishita et al., 2017) を行い，翻訳文の長さを補正した．補正のためのスコアは，式 (2) を使用した．翻訳文の長さを補正することにより，BLEU スコアにおける簡潔ペナルティの影響を受けずに，翻訳文の品質を比較することができるようになる．

逆翻訳も同一のシステムで行った．ただし，生成は 2.1 節の方法で 10 個の疑似原文を生成し，そこからフィルタリングして疑似対訳文を作成した．

**対比方式**　本稿では，ベース対訳のみの場合をベースラインとし，単言語 GCP コーパスを人手逆翻訳したもの（すなわち本来の対訳文）をすべて追加した場合を，翻訳品質の上限と考える．そして，以下の方式の比較を行う．

- 目的言語 1 文あたりの疑似原文数による翻訳品質
- 2.3 節で述べたフィルタリング方法 3 種類
- 疑似原文生成方法を，サンプリングではなく，N ベスト生成にした場合

**評価**　翻訳品質は，BLEU (Papineni et al., 2002) で評価する．検定は，multeval ツール（ブートストラップ再サンプリング方式）[2]を用い，有意水準を 5%として行った（$p < 0.05$）．

## 3.2　GCP コーパスでの実験結果

まず，単言語コーパスとして，GCP コーパスを用いた結果について説明する．

[2] https://github.com/jhclark/multeval

表 3: 単言語コーパスに GCP コーパスを用いた場合 (疑似原文数 6 のとき．カッコ内はベース対訳のみからの増分)

| 方式 | 英日翻訳 BLEU | 中日翻訳 BLEU |
|---|---|---|
| ベース対訳のみ | 26.19 | 37.08 |
| 尤度による選択 | 30.41 (+4.22) | 42.14 (+5.06) |
| 信頼度による選択 | **30.59** (+4.40) | **42.22** (+5.14) |
| ランダム選択 | 30.27 (+4.08) | 42.09 (+5.01) |
| N ベスト生成 | 29.62 (+3.43) | 40.77 (+3.69) |
| 人手逆翻訳 | 31.05 (+4.86) | 42.37 (+5.29) |

GCP コーパスを追加した場合の疑似原文数と BLEU スコアの関係を図 2 に示す．(a) は英日翻訳，(b) は中日翻訳の結果である．また，表 3 は，疑似原文数が 6 のときの数値を取り出したものである．

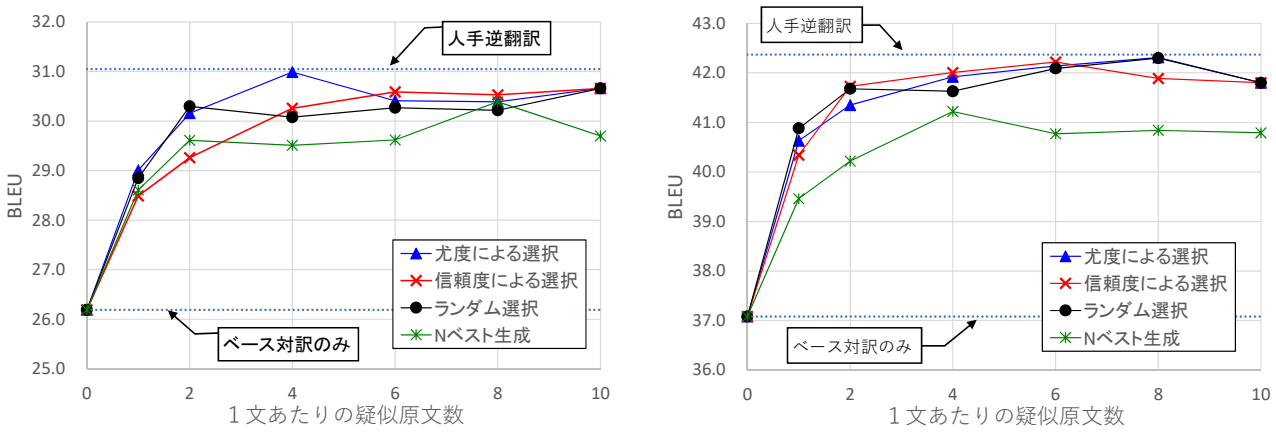
英日翻訳，中日翻訳ともに，単言語コーパス 1 文につき，複数の疑似原文を与えた方が，BLEU スコアは向上する．目的言語の文数は，ベース対訳のみを除き，すべての場合で同じなので，複数の疑似原文を学習することは，エンコーダーの精度向上に有効であると言える．特に，表 3 の尤度，信頼度，ランダム選択のように，ベース対訳のみから人手逆翻訳への BLEU スコア向上分（英日で+4.86）のうち，8 割以上を単言語コーパスで達成できている（同+4.08〜+4.40）点は特筆すべきである．しかし，どの方式も，人手逆翻訳の BLEU スコアには若干届かず，本来の対訳文を代用するところまでは至らない．

次に，疑似対訳のフィルタリング方法（尤度，信頼度，ランダム）を比較すると，データによってばらつきはあるが，ほぼ，どの方式も同じような BLEU スコアを示している．実際，検定を行ったところ，中日翻訳では，すべての場合において有意差はなかった．英日翻訳の場合，いくつかのケースで有意差が現れた（たとえば，疑似原文数 2 のときのランダム選択と信頼度選択）が，有意差なしのケースも多く，強い傾向は見られなかった．また，疑似原文の生成方法を，N ベスト生成にしたところ，提案方式（サンプリング）に比べ，明らかに BLEU スコアが低くなった，

今回，ベース対訳が 40 万文と，元々高品質な逆翻訳器を使用したため，尤度や信頼度によるフィルタリングはあまり効果がなく，疑似原文の多様性の方が精度向上に寄与したものと思われる．

## 3.3　BCCWJ での実験結果

表 4 は，単言語コーパスとして BCCWJ を使用したときの実験結果である．なお，リソースの関係で，




(a) 英日翻訳の場合　　(b) 中日翻訳の場合

図 2: 単言語コーパスに GCP コーパスを用いた場合の疑似原文数と BLEU スコア

表 4: 単言語コーパスに BCCWJ を用いた場合 (尤度フィルタリング)

| 疑似原文数 | BLEU |
|---|---|
| 0 (ベース対訳のみ) | 26.19 |
| 1 | 28.72 |
| 2 | 30.24 |
| 4 | 30.56 |
| (人手逆翻訳) | 31.05 |

今回は英日翻訳，尤度によるフィルタリングのみ，実験を行った．

BCCWJ の場合も，GCP コーパスと同様に，疑似原文数を増加させると，BLEU スコアは向上する．コーパスサイズが異なるので，両者の直接比較はできないが，異なるドメインの単言語コーパスでも，数倍の量を使用すると，同一ドメインコーパスと同程度の品質向上が達成できている．

## 4　まとめ

本稿では，Sennrich et al. (2016a) が提案した，単言語コーパスの逆翻訳法を拡張し，サンプリングによって複数の疑似原文を生成することによって，エンコーダー（およびアテンション）を強化した．疑似対訳コーパス学習の際は，ベース対訳コーパスと学習率を変えて学習して，疑似対訳コーパスに過度に適応する問題を避けた．その結果，目的言語 1 文に対する疑似原文数を増加させると，翻訳品質が向上し，人手逆翻訳に近づくことを確認した．また，より良質な疑似対訳を得るために，疑似原文のフィルタリングを試みた．しかし，今回の実験では，有効性を確認することはできなかった．

今後は，より小規模なベース対訳コーパスでの効果を確認するとともに，原言語側の単言語コーパスの利用も検討してゆきたいと考えている．

## 謝辞

本研究は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として行われました．

## 参考文献

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR 2015*.

Atsushi Fujita and Eiichiro Sumita. 2017. Japanese to English/Chinese/Korean datasets for translation quality estimation and automatic post-editing. In *Proc of WAT2017*, pages 79–88.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. of ACL 2017, System Demonstrations*, pages 67–72.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. NTT neural machine translation systems at WAT 2017. In *Proc. of WAT2017*, pages 89–94.

Yusuke Oda, Katsuhito Sudoh, Satoshi Nakamura, Masao Utiyama, and Eiichiro Sumita. 2017. A simple and strong baseline: NAIST-NICT neural machine translation system for WAT2017 English-Japanese translation task. In *Proc. of WAT2017*, pages 135–139.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL-2002*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. of ACL-2016 (Volume 1: Long Papers)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proc. of ACL-2016 (Volume 1: Long Papers)*, pages 1715–1725.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS 2014*, pages 3104–3112.

今村賢治，隅田英一郎. 2018. グローバルコミュニケーション計画のための多言語パラレルコーパス. **言語処理学会第** *24* **回年次大会**