疑似原文生成によるニューラル機械翻訳への単言語コーパスの導入

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

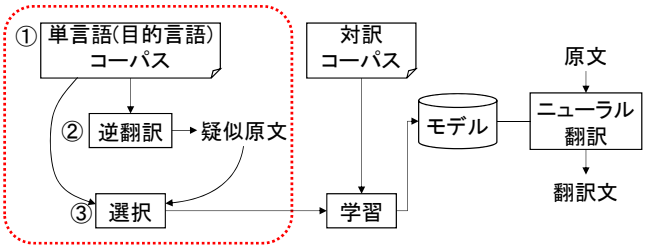

図 1: 本研究の構成

1 はじめに

ニューラル機械翻訳 (NMT) では，近年，sequence-to-sequence 学習 (Sutskever et al., 2014; Bahdanau et al., 2015) が成功を収めてきている．この方法は，原文を実数値から構成される文脈ベクトルに符号化 (encode) し，そこから翻訳文を生成 (decode) する方法である．

ニューラル機械翻訳は，誤差逆伝播によって学習されるため，原文と翻訳文の対，すなわち対訳文が必須となる．一方，フレーズベースに代表される従来の統計翻訳 (SMT) では，対訳文に比べて入手しやすい大規模な単言語コーパスから目的言語の言語モデルを作成し，組み込むことで，翻訳品質を向上させることができる．しかし，単言語コーパスは原文に相当する文を持たないため，ニューラル機械翻訳に組み込むためには，適当な文脈ベクトルを与える必要がある．

Sennrich et al. (2016) は，単言語コーパスを原言語方向に翻訳（逆翻訳）することで疑似原文を生成，これを疑似対訳文として，本来の対訳コーパスとともにニューラル機械翻訳器を学習し，大幅な精度向上が可能であると報告した（図 1）．彼らはデータ追加の際，単言語コーパスからランダムに文を選択したが，どのような文を選択すると翻訳品質に寄与するのか，詳細は明らかにしていない．

本稿では，以下の 2 つの観点から文を選択・追加し，ニューラル機械翻訳における Sennrich et al. (2016) の効果を確認することを目的とする．

- 逆翻訳の翻訳品質による選択．疑似原文が人間の翻訳に近い方が効果が高いと考えられるため，逆翻訳の品質で追加データを取捨選択する．逆翻訳品質はさらに，以下の方法を確認する．
  - 使用する翻訳器を変更（図 1 の ②）
  - 逆翻訳器の翻訳信頼度による文の選択（同 ③）
- 対訳コーパスとのドメイン適合性による選択．単言語コーパスの文（目的言語）および疑似原文と，対訳コーパスの類似性の高さで追加データを取捨選択する．ドメイン適合性は，さらに以下の方法を確認する．
  - 単言語コーパスそのものを変更（図 1①）
  - 言語モデルパープレキシティを利用したドメイン適合度による文選択（同 ③）

以下，第 2 節では，本稿のニューラル機械翻訳の概要と，単言語コーパスの利用法について説明する．第 3 節では，本稿で確認する文の選択方法（逆翻訳品質，ドメイン適合性）について詳細を説明し，第 4 節で実験を行う．なお，本稿では，英日翻訳を対象とする．

2 Sequence-to-Sequence 学習によるニューラル機械翻訳

2.1 アテンション付きエンコーダー・デコーダー

現在のニューラル機械翻訳は，アテンション付きエンコーダー・デコーダー方式 (Bahdanau et al., 2015) が主流となっている．構成を図 2 に示す．

ここでは，原言語の単語列 $(x_1, x_2, \cdots, x_T)$ をエンコーダーによって文脈ベクトル（$\overrightarrow{h_i}$ および $\overleftarrow{h_i}$）に符号化し，そこからデコーダーによって，翻訳文

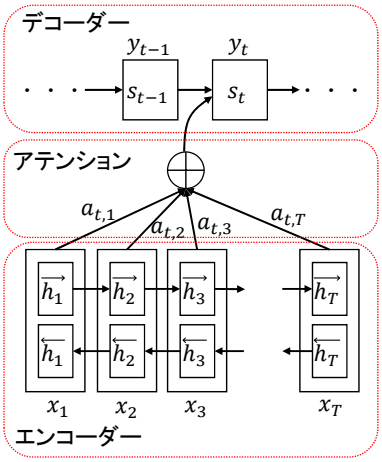

図 2: アテンション付きエンコーダー・デコーダー (Bahdanau et al., 2015)

$(y_1, \ldots, y_t, \ldots)$ を生成する．エンコーダー，デコーダーともに，LSTM(long short-term memory) ベースの再帰ニューラルネットワーク (RNN) で構成されている [1]．エンコーダーで符号化された文脈ベクトルだけでは，どの単語を翻訳したのか不明であるため，アテンション機構によって原言語の単語と目的言語の単語の対応付けを行っている．

このネットワークは，原言語の単語列に対して，目的言語の単語列を正解として与え，誤差逆伝播によって学習する．つまり，学習には，対訳文が必要である．

2.2 単言語コーパスの導入

アテンション付きエンコーダー・デコーダーの学習は，対訳文で行うため，これに単言語コーパスを導入する方法は自明ではない．現在，以下の 2 種類の方式が提案されている．

- **方式 1**: 単言語コーパスだけで RNN 言語モデルを学習し，対訳コーパスから学習したデコーダーの RNN と合成して使用する方法 (Gülçehere et al., 2015)．
- **方式 2**: 単言語コーパス各文の文脈ベクトルを推定して与え，デコーダーの RNN を学習する方法 (Sennrich et al., 2016)．

本稿では，方式 2 を採用する．Sennrich et al. (2016) が提案した方式は，単言語コーパスを，別途対訳コーパスだけから学習した翻訳器で逆翻訳し，疑似原文を得るというものである．この疑似原文をエンコードすると，原言語に対応する文脈ベクトルが得られる．つまり，表面的には疑似対訳文を通常どおり学習するだけで，NMT 自身を変更することなく，単言語コーパスを導入することができる．なお，目的言語は正解であるため，デコーダーのモデルは正確である．Sennrich et al. (2016) では，（言語対や設定により異なるが）方式 1 より翻訳品質が向上できたと報告している．ただし，ニューラル機械翻訳は学習に時間がかかるため，単言語コーパスの文すべてを学習に使うのではなく，対訳コーパスと同サイズ程度にサンプリングして使用する．

[1]LSTM 以外の素子を使うか，双方向/片方向のエンコーダーを使うかなど，実装により異なる場合がある．

表 1: コーパスサイズと追加時の SMT 翻訳品質

種別	コーパス	サイズ (文)		BLEU
対訳	ASPEC	学習	196,165	27.96
		開発	1,790	
		テスト	1,784	
単言語	ASPEC		740,003	29.38*
(日本語)	BCCWJ		4,791,336	28.28*

3 追加データの選択

Sennrich et al. (2016) では，単言語コーパスからランダムに追加データを選択していたが，どのような選択を行うかによって，最終的な翻訳品質は変わると考えられる．本稿では，逆翻訳の翻訳品質と，対訳コーパスとのドメイン適合性の観点でデータを選択する．具体的には，図 1 の 3 箇所のバリエーションを試す．

3.1 単言語コーパスによる差異

今回，対訳コーパスには ASPEC-JE(Nakazawa et al., 2016) のうち，対訳としての対応度が高い約 20 万文を用いる．それに対し，単言語コーパスには，ASPEC-JE のうち，次に対訳対応度が高い約 74 万文 [2]，および現代日本語書き言葉均衡コーパス (BCCWJ)1.1 のうち，1024 文字以下の約 480 万文を使用した．本稿で用いるコーパスと，そのコーパスで学習されたフレーズベース統計翻訳 Moses[3] の翻訳品質を，表 1 に示す．なお，単言語コーパスの BLEU スコアは，対訳コーパスで訓練した翻訳モデルおよび言語モデルに，単言語コーパスの言語モデルを追加したときのスコアである．また，表中の '*' マークは，対訳コーパスのみの BLEU スコアを基準にしたとき，有意差があることを示す（$p < 0.05$）．

単言語コーパス ASPEC は，対訳コーパスとドメインが完全に一致しているため，言語モデルとして追加しても BLEU は大幅に向上する．一方 BCCWJ は，有意差はあるものの，約 480 万文使用したにも関わらず，BLEU スコアの向上は少ない．

[2]80 万文から，50 単語以下の文だけを使用した．
[3]http://www.statmt.org/moses/

3.2 逆翻訳器による差異

単言語コーパスを翻訳する逆翻訳器には，以下の 2 種類を使用する．

- 順方向の翻訳器と同じニューラル機械翻訳器 (NMT)．詳細は 4 節で述べる．
- フレーズベース統計翻訳器 Moses(SMT)．

いずれも学習には表 1 の対訳コーパスのみを用いる．実験では，NMT の BLEU スコアが 29.10 に対して，SMT の BLEU スコアが 27.96 で，NMT の翻訳品質が高い．

3.3 疑似対訳文のスコアによる選択

3.3.1 翻訳信頼度

翻訳器が出力する尤度は，信頼度としても扱われ，翻訳文の品質を表現する一つの指標である．そのため，逆翻訳の尤度が高い文を優先的に追加データとすることによって，効率的に翻訳品質を向上できる可能性がある．ただし，単に尤度が高いだけでは，短い文が選好されてしまうため，単語数で正規化した翻訳信頼度 $score$ を用いる．

$$score = \frac{\text{逆翻訳器が出力した尤度}}{\text{目的言語の単語数}} \quad (1)$$

3.3.2 ドメイン適合度

本稿の目的は，目的言語の単言語コーパスを学習させることであるので，逆翻訳の品質の他に，単言語コーパスと対訳コーパスの適合性が影響する可能性がある．本稿では，この適合性を，言語モデルのドメイン適応に用いられている，エントロピーの差で測定する (Moore and Lewis, 2010)．具体的には，以下の単言語エントロピー差の大きい文を優先して追加データに加える．

$$score = H_{in\_trg}(s) - H_{out\_trg}(s) \quad (2)$$

ただし，$H_{in\_trg}(s)$ は文 s の対訳コーパスの目的言語の言語モデルで測定した交差エントロピー，$H_{out\_trg}(s)$ は単言語コーパスの言語モデルで測定した交差エントロピーである．本スコアが高い場合，その文は対訳コーパスに類似していると見なせる．

4 実験

4.1 実験条件

本稿では，英日翻訳で評価を行う．

使用したニューラル機械翻訳器は，Harvard NMT である [4]．語彙は，表 1 の対訳コーパスに出現する単語のうち，頻度 2 以上のものを使用した（英語 41k, 日本語 35k）．学習には確率的勾配降下法 (SGD) を用い，バッチサイズ 64，学習率 0.1 で 15 エポック，その後半減させながら 5 エポックを実施した．翻訳時には，シングルモデル（アンサンブルなし），ビーム幅 5 で翻訳した．

追加コーパスには表 1 の単言語コーパスを用いたが，ニューラル機械翻訳の語彙において，未知語率が 10%未満の文だけを追加対象とした．

4.2 実験結果

図 3，4 は，単言語コーパスの追加文数に対する，それぞれ BLEU スコア，テストセットパープレキシティを測定したグラフである．また，追加文数 20 万文，40 万文のときの方式別 BLEU スコア，パープレキシティの具体的数値を表 2 に示す．なお，BLEU スコアは翻訳文の品質を直接表し，高い数値が高品質を表している．テストセットパープレキシティは，学習されたモデルがテストセットにどれだけ適合しているかを表しており，低いほどテストセットに適合したモデルであることを示す．以下，表 2 に基づいて議論する．

まず，単言語コーパスによる差異（ランダム追加，逆翻訳 NMT の場合）に着目すると，ベースラインに比べ，ASPEC，BCCWJ ともに単言語コーパス追加によって，BLEU スコアは向上した．しかし，ASPEC の BLEU スコアに比べ，BCCWJ の BLEU スコアは明らかに低い．これは，対訳コーパスとのドメインの差異によって引き起こされたと考えられ，統計翻訳における単言語モデル追加（表 1）と同様な結果が，ニューラル機械翻訳でも起こっている．

次に，逆翻訳器による差異（ランダム追加，ASPEC コーパスの場合）に着目する．SMT の方は，20 万文，40 万文追加によっても BLEU スコアの変動はほぼなかったが，NMT は追加によって有意に向上した．ベースラインシステムでも，BLEU スコアは NMT の方が高かったため，逆翻訳品質が高い方が単言語コーパス追加の効果が高いと言える．しかし，人手による逆翻訳と BLEU スコアの差はほとんどないことを考えると，逆翻訳品質以外の要素が影響している可能性もある．

最後に，疑似対訳文のスコア（ランダム，翻訳信頼度，ドメイン適合度）による差異に着目する．BLEU に関しては，いずれのスコアも大きな違いはない．しかし，パープレキシティに関しては，ランダム追加に

[4] https://github.com/harvardnlp/seq2seq-attn

表 2: 疑似対訳文を追加したときの NMT の BLEU スコア，テストセットパープレキシティ (PPL)
表中の ‘*’ は，ベースラインに対して，有意に向上していることを表す ($p < 0.05$)

方式	逆翻訳器	単言語コーパス	20 万文追加 BLEU	20 万文追加 PPL	40 万文追加 BLEU	40 万文追加 PPL
ベースライン (対訳のみ)	-	なし	29.10	9.818	29.10	9.818
ランダム追加	人手	ASPEC	31.11*	6.597	32.05*	5.850
ランダム追加	SMT	ASPEC	29.16	8.397	28.84	8.138
ランダム追加	NMT	ASPEC	31.70*	8.114	31.91*	7.800
ランダム追加	NMT	BCCWJ	29.41	9.290	29.61*	9.236
翻訳信頼度による選択	NMT	ASPEC	30.99*	8.068	31.97*	7.669
ドメイン適合度による選択	NMT	ASPEC	31.26*	8.015	31.92*	7.643

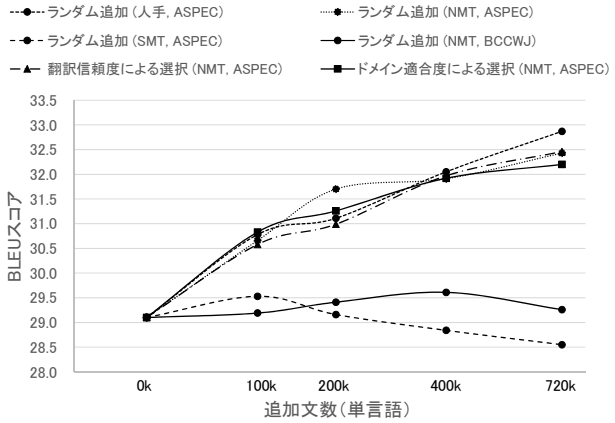

図 3: 追加文数と BLEU スコア

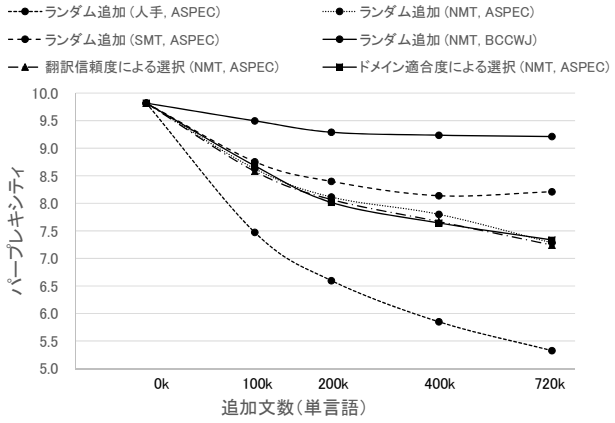

図 4: 追加文数とテストセットパープレキシティ

比べ，翻訳信頼度，ドメイン適合度ともに低下している．たとえば，40 万文追加の場合，ランダム追加のパープレキシティが 7.800 であるのに対し，翻訳信頼度が 7.669，ドメイン適合度が 7.643 である．したがって，翻訳信頼度，ドメイン適合度による選択は，モデルの品質という観点では有効である可能性が高い．

5 まとめ

本稿では，ニューラル機械翻訳に対して，単言語コーパスの追加による翻訳品質の向上について報告した．追加コーパスは，もともとの対訳コーパスと同一ドメインであることが望ましく，疑似原文を生成するための逆翻訳器は，ニューラル機械翻訳器を使用する方がよい．また，逆翻訳された疑似対訳文について，翻訳信頼度およびドメイン適合度で選択を行った．結果として，選択方式による BLEU スコアの向上は認められなかったが，テストセットパープレキシティが低下し，良質なモデルが学習された．

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 *Proceedings of International Conference on Learning Representations (ICLR 2015)*.

Çaglar Gülçehere, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Ling, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. In *CoRR, abs/1503.03535*.

C. Robert Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASEPC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth Edition of the Language Resources and Evaluation Conference (LREC-2016)*, Portoroz, Slovenia, May.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016, Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3104–3112.