# 文脈情報を考慮した日英ニューラル機械翻訳

李 凌寒　　中澤 敏明　　鶴岡 慶雅
東京大学 大学院情報理工学系研究科
{li0123,nakazawa,tsuruoka}@logos.t.u-tokyo.ac.jp

## 1 はじめに

機械翻訳はニューラル機械翻訳の登場により目覚ましい精度の向上を遂げた。一方で、通常の機械翻訳モデルは1文ずつ翻訳する仕組みになっており、ひとまとまりの文書を訳す際に同文書内の他の文の情報を考慮することが出来ない。人間が文書を翻訳をする際には文書全体の情報を考慮しながら、文の意味を正しく解釈し一貫性のある訳語を選択することを考えると、各文の情報を独立に扱う機械翻訳の処理は翻訳の品質に大きな制限を加えている。この問題を解決するために、文脈情報を考慮するニューラル機械翻訳モデルの研究が近年盛んに行われている。

本研究では、訓練・翻訳時に文書内の一つ前の文を考慮するモデルを、日英の新聞記事翻訳に適用した実験の結果を報告する。ベースとなる翻訳モデルは一般的な LSTM モデルを用いた。また文脈を考慮するモデルとして、 Bawden ら [1] の実験を参考にし、通常の対訳文に加えてそれらの文書内での前文を入出力するもの、また文脈情報用のエンコーダーを用意しアテンションに組み入れるものでの実験を行った。時事通信社の新聞記事から作成した日英コーパスを用いた実験の結果、1文前の情報を考慮することによって最大で 0.5 ポイント程度の BLEU [6] 値の上昇が見られた。しかし、ゼロ照応の翻訳などは、1文前を考慮するモデルでは解決することが出来ず、課題は依然として残っている事が示唆された。

## 2 モデル

今回の実験では、1文ずつ翻訳するモデルと、1文前の情報も加えて考慮するモデルの実装、比較を行った。ベースとなる機械翻訳モデルとして、 Luong ら [4] の LSTM を用いたアテンション付きエンコーダー・デコーダーモデルを用いた。文脈付きモデルは、以下に説明する通り、 Bawden ら [1] の実験で使われている手法を実装した。

### データの入力方式

文脈を考慮するための簡単な設定として、モデルの構造は変えずに、入力と出力に前の文（以降、**文脈文**とする）を加えるものが考えられる [1]。この場合2通りの入力方法があり、1つ目では、2文を特殊文字 *@concat@* で繋げて入力し後半の文（**翻訳文**）のみを出力する（以下、**2-to-1** と表記する）。これによりモデルはソース言語側の文脈情報を考慮できる。

2つ目の設定では、2文を入力しそれらの翻訳を全て出力する（**2-to-2**）。これによりモデルは自身の前文に対する出力、つまりターゲット言語側の文脈情報を考慮できる。 2-to-2 の評価時の BLEU 計算時には、翻訳文の翻訳結果のみを比べ、文脈文の出力は無視されることに注意されたい。これら入出力の具体例を表1に示す。

### マルチエンコーダーモデル

マルチエンコーダーモデル [1] では、通常の翻訳文のエンコーダーに加えて文脈文のためのエンコーダーも用意する。文脈文と翻訳文のエンコーダー出力それぞれについてアテンションベクトルを計算し、その2つのベクトルを組み合わせ、通常のアテンションベクトルと同様に扱い計算を進める。モデルの概略を図1に示す。組み合わせる際の手法は以下に挙げる3通りを用いた。

以下、各タイムステップ毎の文脈文について計算したアテンションベクトルを $\boldsymbol{c}^{(1)}$、翻訳文のものを $\boldsymbol{c}^{(2)}$ とし、組み合わせた後のものを $\boldsymbol{c}$ とする。簡単のため、タイムステップを表す添字は省略する。

#### アテンションの連結 (concat)

2つのアテンションベクトルを連結し、元の次元に戻すように線形変換する。

$$\boldsymbol{c} = \boldsymbol{W}_c[\boldsymbol{c}^{(1)}; \boldsymbol{c}^{(2)}] + \boldsymbol{b}_c \quad (1)$$

| 設定 | 入力 | 出力 |
|---|---|---|
| **2-to-1** | The company faced funding difficulties. *@concat@* They left liabilities totaling 25.5 billion yen. | 負債総額は２５５億円。 |
| **2-to-2** | The company faced funding difficulties. *@concat@* They left liabilities totaling 25.5 billion yen. | 会社の資金繰りが悪化。 *@concat@* 負債総額は２５５億円 |
| **1-to-1** | They left liabilities totaling 25.5 billion yen. | 負債総額は２５５億円。 |

表 1: 文の入力設定（1-to-1 は 1 文ずつの通常の入力となる。）

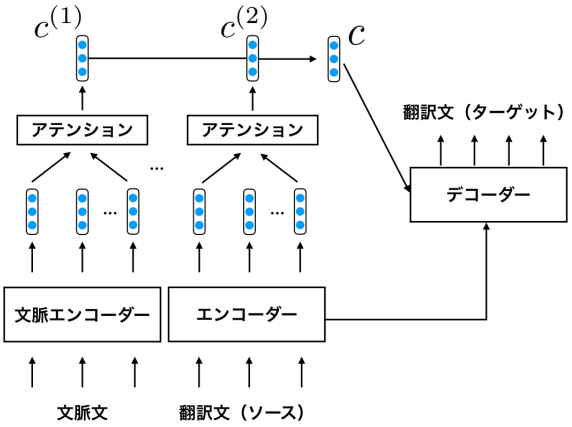


図 1: マルチエンコーダーモデルの概略図

**アテンションの重み付き和 (gate)**

各アテンションに対する重み $r$ を計算し、重み付き和をとる。

$$r = \tanh(\boldsymbol{W}_c[\boldsymbol{c}^{(1)}; \boldsymbol{c}^{(2)}]) + b_r \quad (2)$$

$$\boldsymbol{c} = r \odot \boldsymbol{W}_t \boldsymbol{c}^{(1)} + (1 - r) \odot \boldsymbol{W}_u \boldsymbol{c}^{(2)} \quad (3)$$

**階層式アテンション (hierarchical)**

各アテンションベクトル $\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}$ に関して、再度アテンションを取る形で重みを計算する。以下、$\boldsymbol{z}$ をアテンション計算時の デコーダー LSTM の出力として、

$$e^{(k)} = \boldsymbol{v}_b \tanh(\boldsymbol{W}_b \boldsymbol{z} + \boldsymbol{U}^{(k)} \boldsymbol{c}^{(k)}) + b_e \quad (4)$$

$$\beta^{(k)} = \frac{\exp(e^{(k)})}{\sum_{k'=1}^{2} \exp(e^{(k')})} \quad (5)$$

$$\boldsymbol{c} = \sum_{k=1}^{2} \beta^{(k)} \boldsymbol{U}_c^{(k)} \boldsymbol{c}^{(k)} \quad (6)$$

以上の手法の実験を 2-to-1 と 2-to-2 の設定両方で行った。

# 3 実験

## 3.1 データセット

実験用のコーパスは、時事通信社から提供された日本語と英語の 2011〜2018 年の新聞記事データの本文から、対訳文を抽出したものを用いた。記事データから対訳スコアでフィルタリングする際に文書内の前文にあたる文が取り除かれた翻訳文は、2-to-1, 2-to-2 のモデルであっても 1-to-1 の翻訳をするように学習した。コーパスの統計値を表 2 に示す。

| | train | dev | test |
|---|---|---|---|
| 記事数 | 54,712 | 500 | 500 |
| 文数 | 265,387 (130,008) | 2,362 (1,161) | 2,419 (1,185) |

表 2: 時事コーパスの記事数/文数（括弧内は文脈文を持つ文数）

## 3.2 モデル設定

単語分割は英日共に語彙数を 16,000 に設定した SentencePiece[1] [3] を用いた。機械翻訳のベースとなるモデルは Luong ら [4] のもので、エンコーダー・デコーダーは 2 層の LSTM、単語埋め込みと隠れ層の次元は共に 500 次元、ドロップアウト率 0.3 とした。最適化は Adam [2] を用い、学習率 0.001 から始め、開発データのパープレキシティの最低値が更新されなかったエポック毎に学習率を半減させた。5 回半減したところで訓練を止め、最低値を記録した時点のモデルでテストデータを評価した。

モデルの評価時には、length-normalization [7] を適用したサイズ 6 のビームサーチで翻訳文を出力した。SentencePiece の単語分割は分割を学習するのに用いたコーパスに依存するため、出力文を各言語の単語単

[1] https://github.com/google/sentencepiece

位に変換し（日本語の単語区切りとして KyTea[2] を用いた）、BLEU を計算した。

# 4 結果

## 4.1 定量的評価

各入力設定、モデルを BLEU [6] で評価した結果を表 3 に示す。

| | 1-to-1 | 2-to-1 | 2-to-2 |
|---|---|---|---|
| 英 ->日 | | | |
| - | 16.24 | 16.61 | 15.65 |
| concat | - | **16.78** | 15.77 |
| gate | - | 15.81 | 15.38 |
| hierarchical | - | 10.25 | 8.19 |
| 日 ->英 | | | |
| - | 12.19 | **12.67** | 11.30 |
| concat | | 12.36 | 12.05 |
| gate | - | 12.41 | 12.45 |
| hierarchical | - | 12.42 | 11.71 |

表 3: 各設定の BLEU 値

まず、日 ↔ 英翻訳両方向ともに 2-to-1 の文脈文を考慮するモデルが一番良い BLEU 値を記録している。文脈文を考慮することで、内容や単語の選択に一貫性を持たせた訳が出力しやすくなったためだと思われる。

一方で、2-to-2 の設定では全体的に BLEU 値が下がる結果となった。先行研究 [1] では 2-to-2 の設定による BLEU 値の上昇が報告されているが、そこでは英仏の映画字幕からなるコーパス (OpenSubtitle [3]) が用いられており 1 文の長さが短い。一方で、今回の実験では新聞記事の翻訳をしているため一文の長さが比較的長い。デコーダーが一度に長い文を出力をしなくてはならない場合、出力の途中で不適切な単語を出した際の誤りが伝播していく範囲が増えてしまう。そのため、2-to-2 では出力系列が長くなった分、誤訳も増えたものだと考えられる。

## 4.2 定性的評価：照応の翻訳

1 文のみを考慮するだけでは訳することの出来ない言語現象として照応がある。特に日本語におけるゼロ照応の翻訳の問題は以前から取り組まれてきている [5]。照応先が翻訳文の外にある場合、正しい訳のためには文脈を考慮する必要があるが、この問題は今回の前文を考慮するモデルで解決することは困難であると思われる。その理由として第一に、前文に照応先の情報が含まれているとは限らないという事が挙げられる。例えば、日 → 英翻訳の際に、日本語の原文では省略された主語を英語では補わなければならない状況が多くある。しかし、日本語では主語を一度示した後は複数文に渡って省略することが少なくない。実際にコーパス内の照応表現の参照先を調べると、2 ～ 4 割が 2 文以上前に出現しているということが分かる（図 2）。翻訳時に前文を参照するだけでは、照応を正しく翻訳する情報は得られないことが多いのである（表 4）。

# 5 おわりに

本論文では、1 つ前の文を考慮する文脈付き機械翻訳モデルを、日英の新聞記事翻訳のタスクに適用した実験の結果を報告した。結果として、文脈文を考慮することによる BLEU 値の改善は僅かながらに見られた。しかし、日英翻訳におけるゼロ照応の翻訳などの課題は解決していない。文を超えた単位での翻訳の品質を向上するためには、一つ前の文だけでなく、より広い範囲となる段落や文書全体の情報を考慮して翻訳をするモデルが重要となるだろう。

# 謝辞

本研究の一部は、独立行政法人情報通信研究機構（NICT）の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」の助成を受けて実施された。また図 2 における、OntoNotes のデータ調査は江里口瑛子氏によるものである。深く感謝の意を表したい。

# 参考文献

[1] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of NAACL-HLT*, pp. 1304–1313, 2018.

[2] Diederik P Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*, 2015.

[2] http://www.phontron.com/kytea/index-ja.html
[3] http://opus.nlpl.eu/OpenSubtitles2016.php

| | |
|---|---|
| 入力 | 昨年 11 月には、高知県黒潮町で開かれた「世界津波の日」関連イベントで、30 カ国の高校生ら約 400 人を前に英語でスピーチした。 *@concat@*「忘れることで震災を乗り越えようとしている人たちを傷つけてしまうんじゃないか」と不安もある。 |
| 正解訳 | Still **she** believes it significant to share her experience and the lessons she learned with people from areas that have been free of large-scale natural disasters. |
| 出力訳 | Still **he** said: “It is important to learn from the people’s experiences and lessons learned from the disaster.” |

表 4: 単一エンコーダー 2-to-1 の出力例：ゼロ照応を訳すための手がかりが前の文にあるとは限らない。

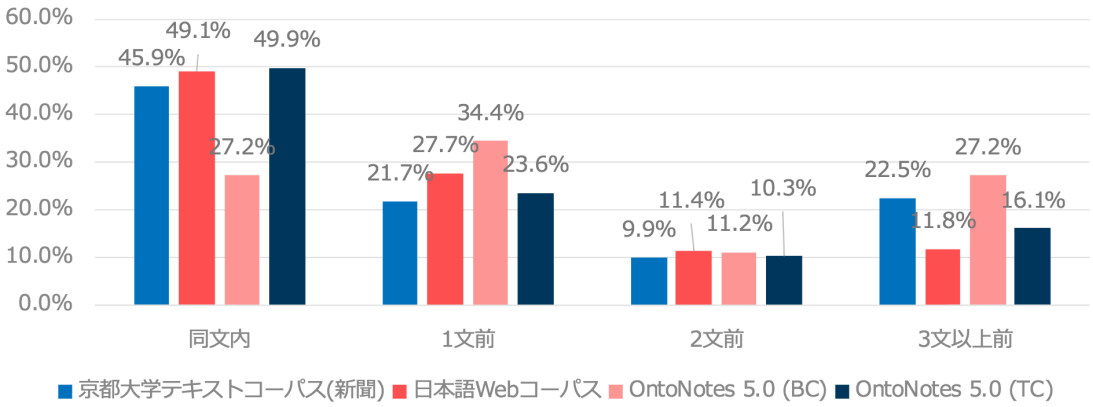


図 2: 照応先が何文前に出現するかを各コーパスで調査したもの（日本語データは萩行らの研究 [8] から抜粋）。日本語データでは省略された要素（ゼロ照応）の先行詞が何文前に出現したか、英語のデータでは全ての共参照関係の先行詞が何文前に出現したかを表す。


[3] Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates . In *Proceedings of ACL*, pp. 66–75, 2018.

[4] Minh-Yhang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP*, pp. 1412–1421, 2015.

[5] Hiromi Nakaiwa and Satoru Ikehara. Zero Pronoun Resolution in a Japanese to English Machine Translation System by using Verbal Semantic Attributes. In *Proceedings of the conference on Applied natural language processing*, pp. 201–208, 1992.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pp. 311–318, 2002.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, and Mohammad Norouzi. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv.org*, 2016.

[8] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–247, 2014.