文脈を考慮した日英機械翻訳に向けた評価データの構築

島津 翔 [1]　高瀬 翔 [1]　中澤 敏明 [2]　岡崎 直観 [1]
[1] 東京工業大学　[2] 東京大学
{sho.shimadu, sho.takase}[at]nlp.c.titech.ac.jp
nakazawa[at]logos.t.u-tokyo.ac.jp　okazaki[at]c.titech.ac.jp

1 はじめに

近年，ニューラルネットを用いた機械翻訳は急速な成長を遂げている [1, 10]．一方，従来の翻訳モデルは 1 文から 1 文への翻訳を行うため，文脈を参照しないと解決できないような一貫性や曖昧性などの問題を考慮しない．図1は一貫性と省略の問題を含む例である．一貫性の例では，文脈文のアトピーを eczema と訳したので，対象文のアトピーも eczema と訳すべきであるが，不正解文では atopy と訳している．省略の例では，対象文からの翻訳としてのみ考えた場合，正解文と不正解文のいずれも正しいが，文脈文を考慮すると正解文が正しい翻訳となる．このように，前後の文で単語の一貫性が取れていない翻訳や，省略された代名詞が正しく補完されない翻訳は，人間にとって非常に読みにくく，文書単位の翻訳としては誤りとなる場合もある．そのため，文脈情報を考慮した翻訳を考えることは重要である．

Bawden ら [2] は，文脈を考慮する翻訳モデルのための評価データを提案した．この評価データには，翻訳対象文よりも前の文脈情報を利用すると翻訳が一意に定まる文が収録されている．この評価データは英仏を対象として構築しているため，日本語など代名詞の言及が必須でない言語に頻出する，省略現象 [8] は扱われていない．

そこで，本論文では日英における，文脈を考慮した翻訳として，代名詞の省略現象に焦点を当てた評価データを構築する．構築した評価データを用いて既存モデルを評価し，省略問題における既存モデルの効果を示す．さらに，翻訳結果の分析を通して，文脈を考慮する既存モデルの問題点を明らかにする．

2 文脈を考慮した評価データ

2.1 評価データの構成

本論文で提案する評価データは，基本的に Bawden ら [2] の提案した評価データと同じ形式を用いる．この評価データは，対象文，文脈文，正解文，不正解文の 4 つで構成されており，対象文は翻訳の対象となる文，文脈文は対象文の直前の文，正解文と不正解文はそれぞれ対象文の翻訳として正しい文と正しくない文を指す．

また，Bawden らは文脈文として対象文の直前の 1 文のみを用いている．これの妥当性を検証するため，後述する OntoNotes [5, 6] を対象に，3 文前までの文間における共参照解析を分析した．結果として，対象文と 1 文前との間に生じる共参照の割合は 44.9% であり，残りの共参照は 2，3 文前との間に生じることから，1 文前だけでは文脈情報として不十分であるとわかった．そこで，本論文では文脈文として対象文の直前の 3 文を利用する．それに伴い，評価データに参照先距離という項目を導入する．参照先距離は何文前の文に必要な文脈情報が含まれているかを示す値である．

- 一貫性
  - 翻訳元
    - 文脈文: 夫は**アトピー**の症状がある.
    - 対象文: 息子も**アトピー**に悩んでいる.
  - 翻訳先
    - 文脈文: Her husband suffered from **eczema**.
    - 正解文: Their son struggled with **eczema**
    - 不正解文: Their son struggled with atopy.
- 省略
  - 翻訳元
    - 文脈文: 彼女は美術館を訪れた.
    - 対象文: そして絵画を鑑賞した.
  - 翻訳先
    - 文脈文: **She** visited museums.
    - 正解文: **She** saw pictures.
    - 不正解文: I saw pictures.

図1: 文外の情報を必要とする例

図2に例を示す．この例は対象文において代名詞が省略されており，対象文だけでは正解文と不正解文のいずれの翻訳も可能であるが，参照先距離が示す文脈文を参照すると，正解文が正しいことが明らかとなる．

2.2 評価データの構築

本論文では，日本語特有の言語現象であり，既存の評価データで扱われてこなかった代名詞の省略に焦点を当て，評価データを構築する．評価データ作成の言語資源として，OntoNotes [5, 6] を用いる．

OntoNotes は，新聞記事やニュース放送など複数のジャンルからなる英語のコーパスであり，人物などのエンティティや共参照情報がアノテートされている．OntoNotes のテキストを英語から日本語に，自然な文となるように翻訳することで，英語側では文をまたがった共参照になっているものが，日本語側では代名詞の省略となる可能性がある．このような文脈を考慮すべき翻訳の事例は対話中に生じやすいと予測されるため，本研究では OntoNotes のニュース放送対話コーパスを利用する．ニュース放送対話コーパスは CNN や MSNBC など 6 つの番組から構成されており，そのうちの CNN 部分を利用し，5,341 文のコーパスから評価データを作成する．

具体的な評価データ作成手順は以下の通りである．まず，OntoNotes の対話文を，男性女性のどちらが話しても日本語の対話文として自然となるように，人手で翻訳する．つぎに，コーパスにアノテートされている共参照チェインを参照し，3 文以内の距離で共参照が生じている文対を抽出する．

そして，抽出した文の日本語訳を人手で確認し，対象文（日本語）において代名詞の省略が生じている文を抜き出す．対象

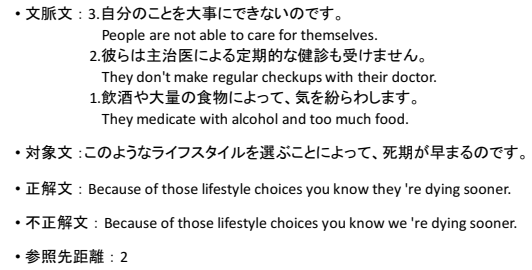

図2: 評価データの例

文（英語）そのものを正解の訳と見なし，対象文（英語）の代名詞をランダムに別の代名詞に置換することで，不正解の訳を作成する．このとき，文内で共参照にある代名詞は個別に置換するのではなく，一貫性が保たれるようにする．さらに，文脈文の日本語訳を参照し，対象文の何文目に文脈情報が含まれているかを確認し，参照先距離を求める．

不正解文は，正解文の代名詞をその他の代名詞にランダムに置換したものであるが，代名詞の置換により動詞の語形を変化させる必要がある場合は，手作業で修正する．また，ランダムに代名詞を置き換えると意味が通らなくなる文や，不正解文として適切でないものが作成される場合がある．例えば，正解文における代名詞が he であり，文脈文から性別が判断できない場合，置き換える代名詞を she とするだけでは，対象文の翻訳として間違った文とならず，不正解文として不適切となる．このような場合には，不正解とならない代名詞を候補から外し，残りの代名詞からランダムに選択する．

図2を例に説明する．最初に対象文と文脈文を自然な日本語に翻訳する．続いて，共参照チェインを調べ，文脈文の 1 文目の they，2 文目の they，3 文目の themselves，正解文中の代名詞 they が共参照チェインを形成していることがわかる．その後，対象文の日本語訳を見ると，正解文の代名詞 they に対応する日本語訳が省略されているのがわかる．そこで，代名詞 they をランダムに選んだ代名詞 we に置換し，不正解文を作成する．最後に，文脈文の日本語を参照すると，文脈文の 2 文目に「彼らは」という文脈情報が見つかるので，参照先距離は 2 とする．以上の手順により，本研究では 506 事例を含む評価データを作成した．その共参照距離の内訳は，距離 1 が 362 件，距離 2 が 96 件，距離 3 が 48 件である．

3 実験

3.1 データセット

日本語から英語への翻訳タスクで実験を行う．本実験の目的は，提案する評価データを用いて既存手法がどの程度省略問題を扱えるか検証し，文脈を考慮する問題に関する翻訳の難しい点を明らかにすることである．

本実験の学習データは，文書単位で翻訳がなされている，すなわち文脈情報を得られる状態の必要がある．そこで，学習データとして OpenSubtitles[*1]から収集した日英字幕コーパスと，時事通信社から提供された日英新聞記事コーパスを用いる[*2]．

OpenSubtitles の字幕データは Bawden のクローリング・前処理スクリプト[*3]を用いて字幕データの収集，前処理を行い，150 万文に対して BPE[9] を適用してサブワードに分割した．

時事通信社の記事データは，日英間で文の対応が 1 文対 1 文ではないため，Bourlon ら [3] の手法を用いて文間の対応スコアを計算し，スコアが 0.3 以上の文を抽出した．そして，得られた 29 万文の記事データに対して BPE を適用し，サブワードに分割した．

いずれの学習データも語彙サイズは 30,000 とした．

3.2 評価方法

翻訳モデルの評価は一般的に BLEU によって行われる．しかしながら，BLEU は正解との，単語の連接の一致数を元にしたスコアであり，文脈を考慮した翻訳を行っているかの評価には適さない．そこで，本研究では構築した評価データを用い，正解文（文脈を考慮した翻訳文）について，モデルのスコアが不正解文より高いか否かの 2 値分類で評価する．

3.3 評価に用いるモデル

本研究で構築する評価データが文内の統計値から解けないと確認するため，多数決モデルと共起モデルという統計値を利用する 2 つのモデルを用いる．

また，現状の翻訳モデルが本研究で構築する評価データをどの程度解けるか比較するため，2 つのニューラル機械翻訳モデルを用いる．

多数決モデル　代名詞の出現頻度を用いて正解文と不正解文のいずれかを選択する．提案する評価データは正解文と不正解文の間で代名詞のみが異なるため，学習データ内の代名詞の出現頻度を計算し，出現頻度が高い代名詞の現れる文を選択する．

共起モデル　代名詞の周りに現れる単語の共起統計から正解文と不正解文のどちらが尤もらしいか推定する．本研究では，共起統計量の近似として，単語ベクトルの内積を用いる．文内の単語を代名詞と代名詞以外に分け，word2vec を用いて単語ベクトルを得る．代名詞の単語ベクトルの集合を $W = \{w_1 ... w_n\}$，代名詞以外の単語ベクトルの集合を $W' = \{w'_1 ... w'_m\}$ として以下の式で値を計算し，正解文と不正解文から値の大きい方を選択する．

$$\sum_{i=1}^{n} \sum_{j=1}^{m} w_i \cdot w'_j \quad (1)$$

評価時に，正解文，不正解文のみを与える手法（Target）と，正解文，不正解文に対象文を連結したものを与える手法（Source+Target）の 2 つで実験を行う．

シングルエンコーダ　文脈を考慮しないニューラル機械翻訳のモデルとして，アテンションを用いた GRU エンコーダ・デコーダモデル [1] を用いる．

Nematus[7] の実装を利用し，ハイパーパラメータは文の最大長を 100 単語とする以外は Bawden ら [2] と同じ値を用いる．

マルチエンコーダ　文脈を考慮したニューラル機械翻訳のモデルとして，Bawden ら [2] はエンコーダを 2 つ用いるマルチエ

[*1] https://www.opensubtitles.org/ja

[*2] ただし，ニューラル機械翻訳モデルの BLEU での評価のため，学習前に各データの 1 部をランダムに抽出し，BLEU の評価データとした．

[*3] https://github.com/rbawden/PrepCorpus-OpenSubs

学習データ	モデル	BLEU
日英字幕	シングルエンコーダ	19.17
	マルチエンコーダ	19.18
日英記事	シングルエンコーダ	17.84
	マルチエンコーダ	16.51

表1: 学習データ別のモデルの BLEU 値

モデル	距離 1 (362 個)	距離 2 (96 個)	距離 3 (48 個)	合計 (506 個)
日英字幕コーパス				
多数決	47.2	49.0	50.0	47.8
共起 (T)	53.0	46.9	75.0	54.0
共起 (S+T)	53.6	46.9	60.4	53.0
シングル	59.4	60.4	66.6	60.3
マルチ	69.6 (0.0013)	55.2 (0.405)	66.6 (1.0)	66.6 (0.0055)
日英記事コーパス				
多数決	49.4	53.1	54.1	50.6
共起 (T)	55.5	51.0	45.8	53.8
共起 (S+T)	58.0	53.1	52.1	56.5
シングル	59.4	57.3	56.3	58.7
マルチ	60.5 (0.777)	58.3 (1.0)	58.3 (1.0)	59.9 (0.677)

表2: 評価データの正解率 (%)．マルチエンコーダの結果における括弧内の値は，シングルエンコーダの出力とマルチエンコーダの出力をマクネマー検定した際の p 値．

ンコーダ・デコーダモデルを提案した．このモデルは，1 つのエンコーダを対象文のエンコード，もう 1 つのエンコーダを文脈情報を持つ文のエンコードに利用し，文外の文脈情報を活用する手法である．

本実験では，2 つのエンコーダにそれぞれ対象文と参照先距離 1 の文脈文を入力し，対象文に対応する翻訳文を出力する 2-to-1 モデルを用いる．

実装は Bawden が公開している Nematus ベースのコードを利用し，ハイパーパラメータは文の最大長を 100 単語とする以外は Bawden ら [2] と同じ値を用いる．

3.4 翻訳全体の質

シングルエンコーダ，マルチエンコーダの翻訳の質として，BLEU スコアを表1に示す．なお，ここでの BLEU スコアは，本研究で構築した評価データではなく，記事・字幕の両データから，学習前にランダムに抽出したデータを用いて計算している．学習データとして日英字幕を用いた場合，シングルエンコーダとマルチエンコーダで値に差はなかった．一方，学習データとして日英記事を用いた場合，マルチエンコーダのほうが 1.3 だけ低下した．これは，マルチエンコーダの出力する文が短い傾向にあるため，BLEU の計算における，短い文へのペナルティが生じるのが原因である．

3.5 結果・分析

表2に，本研究で構築した評価データを用いてモデルを評価した結果を示す．

多数決モデルはどちらの学習データでも 50% 程度の正解率で，ランダムに選択する場合と変わらない結果となった．また，共起モデルはどちらの学習データでも多数決モデルを 3% から

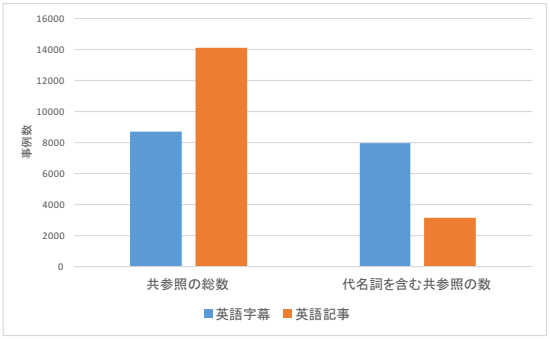

図3: 1 万文あたりの共参照の総数と代名詞を含む共参照の数．対象文と直前 3 文の間に生じる共参照が対象

6% 程度上回っている．シングルエンコーダは，共起モデルより更に正解率が向上している．しかしながら，合計における正解率はシングルエンコーダでも 60% 程度と低く，構築した評価データが代名詞の出現頻度や代名詞周辺の単語の共起統計のみでは正解できないことがわかる．すなわち，構築した評価データは文内の情報や 1 文対 1 文の翻訳では正解できないデータとなっている．

学習データとして日英字幕を用いた場合，マルチエンコーダはシングルエンコーダと比較して参照先距離 1 の正解率が 10% 以上上がっており，本研究で構築した評価データに対して現状の文脈を考慮したモデルとして一番高い値となってる．また，有意水準を $\alpha = 0.01$ とするとマクネマー検定の結果も有意である．したがって，マルチエンコーダは参照先距離 1 の問題に対しては文脈情報を用いて省略された代名詞を当てられていると考えられる．

一方で，参照先距離 2，3 においては，シングルエンコーダとマルチエンコーダの出力の間に有意差はない．これは，マルチエンコーダが 1 文前までしか考慮しないモデルのために，参照先距離 2，3 の問題に効果が無かったためと考えられる．そのため，参照先距離 2，3 の問題にも対応するには，2，3 文前をエンコードするなど，より広い文脈情報を利用する必要があるとわかった．

学習データとして日英記事を用いた場合，全ての参照先距離において，シングルエンコーダとマルチエンコーダの間に大きな差はなく，マルチエンコーダが文脈を考慮した翻訳をできているとは言えない結果が得られた．原因として，日英記事コーパス内に代名詞を含む共参照が少なく，代名詞の省略問題に対する学習が不十分な可能性が挙げられる．図3に Stanford CoreNLP[4] を用いて各学習データの英語側の共参照数を解析した結果を示す．1 万文あたりに生じる共参照の総数では英語記事のほうが多いが，対象文に代名詞を含む共参照に限定した場合，英語字幕の方が多くなる．言い換えれば，英語記事においては，代名詞への言い換えが起きづらい．このため，学習データに含まれる代名詞の省略問題のサンプル数は日英記事の方が少ないと予測され，学習が不十分であったと考えられる．

図4にニューラル機械翻訳モデルの出力を示す．各出力は構築した評価データから各モデルが選択した事例ではなく，対象文の翻訳結果である．1，2 行目の例では，シングルエンコーダ

文脈文	対象文	正解文	シングルエンコーダ	マルチエンコーダ
彼女はまだ俳優業を？	俳優ではありません。	She's not an actor.	It's not an actor.	She's not an actor.
クリントン大統領の下で国家安全保障大統領補佐官を務めたサンディ・バーガー氏についてはどうでしょう。	会談に出席していましたね。	He was at the meeting.	We were in the conference room.	He was at the meeting.
私たちは、ただ食べるためのキャンディそのものを扱うだけではありません。	チョコレートの香りの入浴剤やバニラのボディ用保湿剤のようなスパ製品もあります。	We have a spa product like chocolate bath powders and vanilla body moisturizers.	I've got a product, like, some of the chocolate stuff, or a body.	We've got a product like a diet of chocolate and a couple of bodies.
えー、上院外交委員会の民主党有力者であるジョー・バイデン氏は少し前にイラクにいました。	今週ワシントンで演説するために戻ってきました。	He came back gave a speech in Washington this week.	I came back to make a speech in Washington this week.	He's back in DC this week.
ええと、ウィリアムズは有罪判決を受けた殺人犯です。	しかし国中の人々が命を救おうとしています。	But people all over the country are trying to save his life.	But people in the country are gonna save lives.	It's about the reason he's trying to save lives.
ハロウィーンが頭に浮かぶと、私たちが何を考えたかお分かりですね。	ウィリー・ウォンカやチョコレート工場のことを考えました。	We thought of Willy Wonka and the Chocolate Factory.	I was thinking of Willy Wonka and the chocolate factory.	I thought about the chocolate factory or the chocolate factory.

図4: 日英字幕を用いて学習したニューラル機械翻訳モデルの出力．文脈文は参照先距離 1 の文を表示

は代名詞を間違っているが，マルチエンコーダは下線部が示すように正しい代名詞を予測して正解文を出力している．一方で，3 から 6 行目の例が示すように，マルチエンコーダはシングルエンコーダと比較して出力が大きく異なり，翻訳全体の質が低下する場合がある．具体的には，無関係な単語が出現する（3 行目），必要な情報が抜ける（4 行目），意味の異なる文になる（5 行目），同じ単語を繰り返す（6 行目）といった事例がある．そこで，シングルエンコーダとマルチエンコーダの出力がどの程度異なるか調べるため，単語の一致率を計算すると 51.3% であった．つまり，シングルエンコーダとマルチエンコーダの出力は 50% 近くの単語が異なる，または消失している．この中にはシングルエンコーダで正しく翻訳されている文がマルチエンコーダでは誤った翻訳になる事例（図4，4 行目，5 行目）もある．これは，マルチエンコーダが文脈文と対象文からの情報の取捨選択に失敗しているためと考えられ，改善が必要である．

4 おわりに

本論文では，日本語から英語への文脈を考慮した翻訳として，代名詞の省略現象に着目し，評価データを構築した．統計値を利用したモデルを評価した結果，構築した評価データは文脈を用いなければ正しい回答を行えないことがわかった．ニューラル機械翻訳モデルを用いた実験より，参照先距離 1 の問題に対して，1 文前の文を考慮するモデルが高い正解率を挙げ，文外の情報を用いて省略されている代名詞を当てられると示した．一方で，参照先距離 2，3 の問題において，既存の 1 文前のみを考慮するモデルは文脈を考慮しないモデルとの間に有意差はなく，文脈を考慮する問題を解くにはより広い範囲の文脈情報を活用する必要があると示した．また，ニューラル機械翻訳モデルの出力の分析より，文脈を考慮するモデルは文脈文を用いると，文脈を考慮しないモデルと比較して，翻訳全体の質が低下する場合があるとわかった．

今後は，Bawden ら [2] の構築した英仏データにあるような，一貫性や曖昧性に関する評価データを日英でも作成する予定である．また，文脈情報の取捨選択をより正確に行えるようなモデルを考えたい．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究により得られたものです．

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *ICLR* (2015), pp. 1–15.
[2] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. “Evaluating Discourse Phenomena in Neural Machine Translation”. In: *NAACL* (2018), pp. 1304–1313.
[3] Antoine Bourlon, Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. “Simultaneous Sentence Boundary Detection and Alignment with Pivot-based Machine Translation Generated Lexicons”. In: *LERC* (2016).
[4] Christopher D Manning, John Bauer, Jenny Finkel, and Steven J Bethard. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *ACL* (2014), pp. 55–60.
[5] Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. “OntoNotes: A Unified Relational Semantic Representation”. In: *ICSC* (2007), pp. 517–524.
[6] Weischedel Ralph, Palmer Martha, Marcus Mitchell, Hovy Eduard, Pradhan Sameer, Ramshaw Lance, Xue Nianwen, Taylor Ann, Kaufman Jeff, Franchini Michelle, El-Bachouti Mohammed, Belvin Robert, and Houston Ann. “OntoNotes Release 5.0 LDC2013T19”. In: *Linguistic Data Consortium* (2013).
[7] Sennrich Rico, Firat Orhan, Cho Kyunghyun, Birch Alexandra, Haddow Barry, Hitschler Julian, Junczys-Dowmunt Marcin, Läubli Samuel, Valerio Antonio, Barone Miceli, Mokry Jozef, and Nadejde Maria. “Nematus: a Toolkit for Neural Machine Translation”. In: *EACL* (2017), pp. 65–68.
[8] Sasano Ryohei and Kurohashi Sadao. “A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames”. In: *IJCNLP* (2011), pp. 758–766.
[9] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*. 2016, pp. 1715–1725.
[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need”. In: *NIPS* (2017).