ゼロ代名詞データ拡張による日英機械翻訳の改善

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

ニューラル機械翻訳は単文単位の翻訳において高い性能を発揮する一方で、談話単位での文脈が関わる言語現象を扱う点において未だ課題が残っている。その課題の一つが日英翻訳における**ゼロ代名詞**の翻訳である。日本語では文脈から聞き手が推定可能な主語や目的語を省略することができるが、これを英語に翻訳する際は省略された項を推定した上で適切に翻訳しなければならない。例えば、以下の例文では日本語で省略された主語が一人称であることを推定して、英語では *I* を出力しなければならない。

うなぎが食べたいな。
I feel like eating eel.

ゼロ代名詞の推定には、本質的には談話内におけるトピック、旧情報、世界知識などを参照して可能になるが、一方でゼロ代名詞を含む文内の言語学的な情報が手がかりとなる場合もある [1]。例えば、上記の文では “たいな” という願望を表す機能語列が、主語が一人称であることを示唆している。

この局所的文脈情報とゼロ代名詞の対応を学習することは、既存の単文単位でのニューラル機械翻訳でも解決できるはずの問題であるが、低リソースの状況下ではその対応を十分に学習できない可能性がある。特に、ゼロ代名詞を多く含む傾向にある会話ドメインの翻訳は、現状パラレルコーパスが豊富だと言い難い状況にある。

従って、本研究では局所的文脈情報とゼロ代名詞、特に人称代名詞の翻訳の対応の学習を促進する手法として、ゼロ代名詞データ拡張を提案する（図 1）。本手法では、訓練データを対象に翻訳元言語に現れる人称代名詞を削除し、人工的にゼロ代名詞を含むような翻訳対を作り出す。これにより局所的文脈情報とゼロ代名詞の対応を学習できるような訓練データを増やすねらいがある。

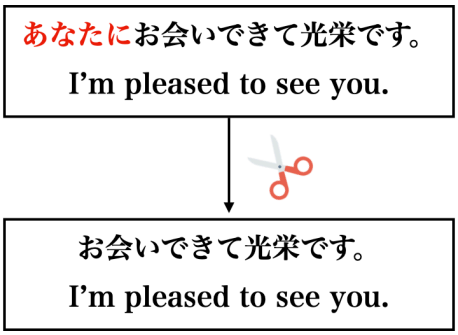

図 1　ゼロ代名詞データ拡張。

本論文では、まず局所的文脈情報からある程度ゼロ代名詞を予測できることを示した上で、実際にゼロ代名詞データ拡張によって機械翻訳モデルのゼロ代名詞予測の精度が上がることを示す。

2 関連研究

2.1 文脈つきニューラル機械翻訳

単文単位の翻訳品質が向上するのに従って、その範囲で扱えないような文脈を考慮した翻訳をするモデルの研究が近年盛んに行われている。最もシンプルなアプローチとして、モデルの構造は変えず、入力や出力を複数文に拡張した 2to1, 2to2 モデルが提案されており、文脈情報を捉えた翻訳の品質を向上させることが示されている [2]。

2.2 ゼロ代名詞

日本語を始めとする一部の言語では述語の項が省略されることがあり、省略された項をゼロ代名詞 (zero pronoun) と呼ぶ。この省略された項を同定するタスクはゼロ照応解析と呼ばれ以前から研究されているが [3]、本研究では特に翻訳におけるゼロ代名詞の問題について考える。

翻訳の際に問題となるのが、ゼロ代名詞が翻訳先の言語では統語的に必須な要素である場合である。

この場合、翻訳の際に省略されたゼロ代名詞を補って翻訳する必要があり、これには翻訳元の文章に加えて、その前の文章などの広い文脈情報が必要となる。

一方で、文内に存在する局所的文脈情報が、決定的でないにしろ、手がかりになる場合もある [1, 4]。例えば、日本語では敬語表現（「申し上げる」）や依頼表現（「～してくれますか？」）などがゼロ代名詞の推測に有用である。

本研究は、広い文脈情報を明示的にモデル化したり、ゼロ照応解析を明示的に行うことをせず、局所的文脈情報から推測できるゼロ代名詞の翻訳を向上させることを主なねらいとする。同様の問題意識に基づく研究として、Wang らによる対訳データのゼロ代名詞を補ったデータを作成して訓練に使うものがある [5, 6]。しかし、Wang らの手法では、対訳データからゼロ代名詞を補うときに英文と中国語文の単語アラインメントが対角線上に並ぶことが多いというヒューリスティクスを用いているが、この手法は語順が大きく異なる日英翻訳に適用することが難しい。そこで本研究では、ゼロ代名詞を補ったデータを活用するのではなく、元々代名詞を含む文からその代名詞を削除したデータを活用する、ゼロ代名詞データ拡張を提案する。

2.3 データ拡張

データ拡張 (data augmentation) は画像分野で盛んに行われており [7]、訓練データに対してタスクにおける意味を損なわない範囲での処理を施すことでデータ量を増やす手法のことである。自然言語処理の分野においては、入力文中の単語を同義語に置換したり [8]、折り返し翻訳により同じ意味を持つ文を生成することで [9]、データ拡張する手法が提案されている。

本研究では、日英翻訳において、翻訳元の日本語の代名詞を削除することで、翻訳元の文の意味を保存したままゼロ代名詞を含む文を獲得する。

3 ゼロ代名詞と局所的文脈情報の分析

本研究は和文内の局所的文脈情報が、英文へと翻訳する際のゼロ代名詞の推定に有用であるという仮定に基づいている。そこで本節では、実際に局所的文脈情報からどの程度ゼロ代名詞を推定できるのか、またどのような局所的文脈情報が有用であるのかを分析する。

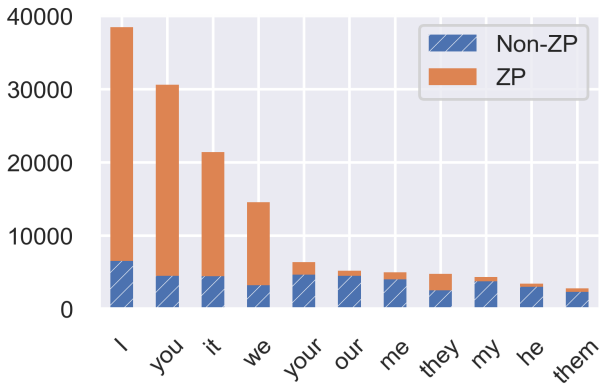

図 2 分析データ中の英語代名詞について、翻訳が和文に現れるものと (Non-ZP) 現れないもの (ZP) の数。

分析データとして日英ビジネスシーン対話コーパス [10] から公開されている対訳データに加えて、研究グループ内で使用できるものを合わせて 104,961 文対を用いた。

3.1 ゼロ代名詞を含む対訳を同定する

和文におけるゼロ代名詞の同定には日本語の構文解析器を用いる方法が考えられるが、今回は翻訳に関わるゼロ代名詞を分析するため対訳コーパスからのアラインメントを活用した方法を採用する。具体的な手順は以下の通りである。

1. 分析用の対訳データから GIZA++[1] を用いて単語アラインメントを学習する。単語分割には、和文に Mecab[2]、英文に spaCy[3] を用いた。
2. 英文中の代名詞が和文の NULL に対応づけられているとき、その英文での代名詞が和文ではゼロ代名詞として実現されているとみなす。

この結果、得られた代名詞の数を図 2 に示す。会話ドメインにおいては、一人称代名詞 *I* や二人称代名詞 *you* が頻出かつ、日本語では省略されやすいことがわかる。また、基本的に英文では高頻度の代名詞ほど、日本語ではゼロ代名詞になることが示されている。

3.2 ゼロ代名詞と共起している局所的文脈情報を抽出する

検出された和文のゼロ代名詞を特徴付けるために、その述語部に現れる単語を抽出する。今回は日本語の構文解析器を用いておらず、ゼロ代名詞は英

1) https://github.com/moses-smt/giza-pp
2) https://taku910.github.io/mecab/
3) https://spacy.io/

文の代名詞とアラインメントで結び付けられているため、英文の係り受け構造をアラインメントを通じて和文に投影することで述語部を抽出する。前項までの処理に引き続き、以下の手順で行った。

3. ゼロ代名詞に対応している英代名詞の係り受け先の単語を抽出する。
4. その単語にアラインメントで対応づけられている和文中の単語と後続の機能語列[4)]を局所的文脈情報として抽出する。

3.3 局所的文脈情報からゼロ代名詞を予測する

この局所的文脈情報から、ゼロ代名詞がどの程度推定可能かを調べるために、ロジスティック回帰を用いた分析を行った。局所的文脈情報として抽出した単語列からのユニグラム、バイグラム、トライグラムを入力素性とし、英文における代名詞を予測する。５分割交差検証による各代名詞の再現率を表 1 に示す。ベースラインとして、訓練データに含まれる代名詞の分布に従ってランダムに予測した場合の値を採用している。

頻度の高い *I, you, we* は、局所的文脈情報を用いた場合ベースラインに比べて有意に高い精度で予測できることが分かる。しかし一方で、頻度の低いその他の代名詞はベースラインと比べても同等か低い数値を記録している。これら頻度の低い代名詞は、局所的文脈情報からだけでは予測することが困難であることが示唆される。

具体的にどのような局所的文脈情報が予測に有用なのかを調べるために、ロジスティック回帰の各出力ラベルに対して、入力素性に対応する重みの値が高いものを抽出し、解釈可能なもの調べた。その結果、一人称単数 *I* の予測には、認識に関わる動詞（思う、わかる、感じる）、謙譲語（申し上げる、存じる）、間投助詞（なぁ、よ）、願望を表す助動詞（たい）、二人称単数 *you* については、質問を表すフレーズ（かな？、ました？）、推測（でしょ、だろ？）、尊敬語（仰る、いただける）、一人称複数 *we* には義務（なきゃ、べき）、願望（たい）が有用であると分かった。その他の代名詞については、代名詞予測に有用であると解釈可能なものは認められなかった。

4） ここでは機能語を Mecab で定義されている品詞の ["助詞", "助動詞", "記号"] とした。

4 実験

前章の分析により、局所的文脈情報がゼロ代名詞の予測に有用であることが確認できた。ここでは、翻訳元である和文中の省略可能な代名詞を助詞とともに削除することでゼロ代名詞を含むデータを人工的に作り出す、ゼロ代名詞データ拡張が有用であるかどうかを検証する。こうして作られたデータはゼロ代名詞と局所的文脈情報の結びつきを学習するために有用な訓練信号を与えると考えられる。

人工データの使用法として、訓練データにそのまま加える、もしくはゼロ代名詞補完をマルチタスクで学習させる手法 [6] が考えられる。本論文では、訓練データに加える方法を検討する。

4.1 実験設定

ゼロ代名詞データ拡張

本研究におけるゼロ代名詞データ拡張では、翻訳元である日本語の文中の省略可能な代名詞を助詞とともに削除することで、ゼロ代名詞を含むデータを人工的に作り出す。代名詞の検出には、人手でパターンを作成し該当する文字列を検索することで行った。パターンの具体的な作成方法は付録を参照されたい。

コーパス

今回の実験は日英会話コーパス [11] を公開されているものに加え、プロジェクト内で使用可能分を加えたものを用いた。データの統計を表 2 に示す。

モデル

翻訳モデルは Transformer [12] を用いた。モデルサイズは Vaswani らの base モデルを 4 層に減らしたもの、最適化器には Adam を用いた。今回は、通常の単文単位の翻訳の他に、入力に前文を付加した 2to1 の設定 [13] についても実験を行った。

評価方法

日英会話コーパスにおける BLEU [14] と日英翻訳ゼロ代名詞評価データセットを用いて評価した [15]。ゼロ代名詞評価データセットは、入力文、正しい代名詞を含む出力文、誤った代名詞を含む出力文のデータからなる。モデルの評価には、モデルに入力文と出力文を入力し、正しい出力文に対する

	I	you	we	they	he	she	us	them	him	her
baseline	35.9	25.4	11.0	3.7	2.2	0.0	2.2	1.9	1.2	0.9
logistic regression	78.2	46.3	17.3	3.8	3.1	0.0	3.6	0.2	0.2	2.9

表 1　代名詞毎のゼロ代名詞予測の再現率。

	train	train+pro_aug	dev	test
文数	246,541	282,952	2,051	2,020

表 2　日英会話コーパスの文数

パープレキシティが低い、つまり正しい出力文の方を予測する確率が高い場合に正答したとみなして精度を算出した。

モデル選択

ニューラルネットワークに基づくモデルはハイパーパラメータやランダムシードの選択に大きく結果が左右されることが知られている。今回は、各設定において一定の計算資源を投入した際の一番良いモデルを用いて性能を比較した [16]。具体的には、ハイパーパラメータとして学習率、ドロップアウト率、ラベル平滑化の値を Optuna[5] を用いて探索し、10 回の試行の内、開発データにおける BLEU スコアが一番高いものについて、テストデータの BLEU とゼロ代名詞評価の精度を算出した。より詳細な設定については付録を参照されたい。

4.2　実験結果

実験結果を図 3 に示す。

	1to1	2to1
baseline	18.2 / 80.5	17.5 / 84.4
baseline+pro_aug	17.7 / 91.1	17.3 / 85.2

表 3　ベースラインとゼロ代名詞データ拡張を行ったモデルの評価。表中の値はそれぞれ BLEU / ゼロ代名詞評価の精度。

1to1 と 2to1 どちらの設定においても、ゼロ代名詞データ拡張によって、BLEU スコアの改善は観察されないが、ゼロ代名詞評価の精度は有意に向上している。1to1 と前文を考慮したモデルである 2to1 を比較すると、baseline の設定においては、前文を加えることでゼロ代名詞翻訳の性能は 80.5% から 84.4% に向上している。一方で、baseline+pro_aug の設定では 91.1% から 85.2% と、前文の文脈を考慮しない方が性能が高い。これは、入力として前文を加える 2to1 の設定では、入力文が長くなるために、その文ゼロ代名詞と局所的文脈情報の関連をモデルが学習しづらくなっているのではないかと推測する。

5）https://preferred.jp/ja/projects/optuna/

5　おわりに

日英の会話翻訳におけるゼロ代名詞翻訳の問題について、和文中のゼロ代名詞を同じ文内に存在する局所的文脈情報からある程度予測できることを示し、この局所的文脈情報とゼロ代名詞の結びつきを翻訳モデルに学習させるための手法としてゼロ代名詞データ拡張を提案した。ニューラル機械翻訳モデルを訓練した実験の結果、ゼロ代名詞データ拡張はゼロ代名詞翻訳の精度を有意に向上させることが示された。

一方で、ゼロ代名詞データ拡張はゼロ代名詞を含む文の外に、ゼロ代名詞翻訳に必要な情報が存在する場合を本質的に解決するものではない。また、分析の結果、局所的文脈情報は一人称、二人称ゼロ代名詞の推測には有用であるが、三人称代名詞の推測には有用ではないことが示唆されており、本手法による改善は限られていると考えられる。これらの課題は、会話中のトピックや登場人物などを捉えることのできるモデルが必要であると考える。

謝辞

本研究成果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」により得られたものです。

参考文献

[1] 工藤拓, 市川宙, 賀沢秀人. Language independent null subject prediction for statistical machine translation. 言語処理学会 第 21 回年次大会 発表論文集, 2015.

[2] Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019.

[3] Kenji Imamura, Kuniko Saito, and Tomoko Izumi. Dis-

criminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, August 2009.

[4] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Japanese zero reference resolution considering exophora and author/reader mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, October 2013.

[5] Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. A novel approach to dropped pronoun translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, June 2016.

[6] Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. One model to learn both: Zero pronoun prediction and translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 2019.

[7] Patrice Simard, Yann LeCun, John S. Denker, and Bernard Victorri. Transformation invariance in pattern recognition-tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, Berlin, Heidelberg, 1998. Springer-Verlag.

[8] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, June 2018.

[9] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, Vol. abs/1804.09541, , 2018.

[10] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong, China, November 2019.

[11] Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. Document-aligned japanese-english conversation parallel corpus. In *Proceedings of the Fifth Conference on Machine Translation*, Online, November 2020.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[13] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, September 2017.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002.

[15] Sho Shimazu, Sho Takase, Toshiaki Nakazawa, and Naoaki Okazaki. Evaluation dataset for zero pronoun in Japanese to English translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020. European Language Resources Association.

[16] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, November 2019.

付録

ゼロ代名詞削除に用いた代名詞+助詞リスト

ゼロ代名詞削除は、代名詞リスト（表 4）と助詞のリスト（表 5）からの全組み合わせを列挙し、そのパターンに該当する文字列を文中から削除することで行った。

以下の代名詞は、コーパスの文を Mecab を用いて品詞解析したときに代名詞にあたる単語から、人手で人称代名詞を抽出したものである。

一人称単数	私, わたし, 僕, ぼく, 俺, おれ, わたくし, オレ, ウチ
一人称複数	我々, 僕ら, われわれ, 僕達, 僕たち, 私達
二人称単数	貴方, 貴女, あなた, お前, おまえ, 君, あんた
二人称複数	君たち, みなさま
三人称単数	彼, 彼女, あいつ
三人称複数	彼ら, 彼女ら, みんな, 皆, 皆んな, みなさん, 奴ら

表 4　ゼロ代名詞削除に用いた代名詞のリスト

助詞については、人称代名詞に続く助詞の中から頻出かつ削除しても文意が伝わると思われるものを人手で選定した。

主格	は , が ,
対格	を
与格	に
所有格	の
その他	も , の方から , のほうから , の方に , のほうに , の方で , のこと , の事 , のほうで , から , 、

表 5　ゼロ代名詞削除に用いた助詞のリスト

機械翻訳実験のハイパーパラメータ

機械翻訳実験におけるハイパーパラメータの値は表 6 に示す範囲で Optuna を用いて探索した。

	low	high
source_embedding_dropout	0.1	0.6
target_embedding_dropout	0.1	0.6
encoder_dropout	0.1	0.6
decoder_dropout	0.1	0.6
label_smoothing	0.1	0.6
lr	0.0001	0.001

表 6　ハイパーパラメータ探索の範囲