文の持つ情報量を用いたニューラル機械翻訳の訳抜け検出

藤井 真　　新納 浩幸　　古宮 嘉那子
茨城大学 工学部 情報工学科
{15t4057f, hiroyuki.shinnou.0828, kanako.komiya.nlp}@vc.ibaraki.ac.jp

1 はじめに

近年、ニューラルネットワークを利用した技術は様々な領域の問題解決に用いられている。自然言語処理の分野ではニューラル機械翻訳 (Neural Machine Translation;NMT) が挙げられる。NMT 以前の機械翻訳は機械が翻訳していることを感じることが多かったが、NMT は人が受け入れやすい自然な文章をもたらしている。

しかし、NMT による翻訳は従来手法である統計的機械翻訳 (Statistical Machine Translation;SMT) において問題となりにくい、訳抜けという問題が起こりやすい [3]。

訳抜けとは、原文に存在していた文意を訳文において意味的、単語的に欠いてしまう現象をいう。SMT では、原文の部分的置き換えが統計的に不都合なく全体に及んだ時点で翻訳終了となるため、訳抜けのある状態で翻訳を終了することは少なかった。しかし、NMT は自然な文とするための流れを重視する傾向があるため、ニューラルネットワークで学習した流れの中に原文の流れが無かった場合など、その部分を抜かして文意を欠いたまま文尾まで流れを作り終え、翻訳を終了してしまうことがある。このような学習外のために翻訳を飛ばした部分が訳抜けとなる。訳抜けの規模としては、1 単語から句、文の半分以上を占める節まで抜ける場合がある。

自然な文をもたらすために翻訳で文意を欠くことは、意思伝達の上で好ましくない。NMT において訳抜けを検出することは NMT に再検討を促し、より文意を含ませた訳文を出力させうる点で有用である。また、原文と訳文の持つ情報量を示すことは、訳文言語を全く知らない被翻訳者に対して原文情報含有率などを提供でき、訳文の言葉足らずや誤解可能性を認知させ、注意喚起を行える点でも有用である。

本論文ではニューラル機械翻訳の訳抜け検出に対して、原文の情報量ベースの手法を提案する。原文の単語が持つ情報量を元に文全体のあるべき情報量を仮定し、同様に仮定した訳文の情報量と比較することで訳抜けを検出するものである。この手法の特徴は、入力として与える情報が原文と訳文の対であることである。本研究は NMT のモデルを対象としているが、SMT のモデルや人が訳した文であっても理論上は訳抜けの検出を行える。また、逆翻訳や NMT の内部へ干渉しないこと、一般的なニューラルネットワークや機械学習が生成するモデルを要しないことなど、低コストに検出を行える。

実験では、英語で書かれたニュースの文を原文とし、NMT により訳文を作成する。それら原文と訳文から提案手法によって訳抜け状態の文を検出し、その結果や例を示す。

2 関連研究

NMT の訳抜けした内容の検出については、ニューラルネットワークの累積アテンション確率と逆翻訳による文生成の確率を利用した手法 [1] がある。ニューラルネットワークの累積アテンション確率による手法は、ニューラルネットワーク内で出力までの評価に用いられるアテンション確率の高低で翻訳された内容か否かを判断する手法である。逆翻訳による文生成の確率を利用した手法は、原文から NMT によって出力した訳文をもう一度 NMT によって原文言語に戻し、原文言語同士の比較によって訳抜けした内容を検出する手法である。結果として後者が優れているとのことだった。前者はニューラルネットワーク内部のアテンション確率と訳文、後者は原文と、二重の NMT によって原文言語に戻した二重訳文を用いている。本研究は原文と 1 度の NMT の訳文だけを用いて、NN モデルに関与せず訳抜けの有無を検出するものである。

また、SMT との併用により NMT の大規模語彙対応と訳抜けを削減させる手法 [2] がある。NMT の弱点である大規模語彙を含む翻訳への対処として SMT

を利用している。その中で NMT の訳抜けの問題が生じた対処を後藤らの逆翻訳確率を利用した手法によって行うものである。この研究は大規模語彙に対応し、かつ訳抜けを削減した NMT モデルを作成する手法であり本研究の訳抜けを検出する手法とは異なる。

3 提案手法

本研究は原文の単語が持つ情報量を元に文全体のあるべき情報量を仮定し、同様に仮定した訳文の情報量と比較することで訳抜けを検出するものである。図 1 は提案手法の概要図である。原文から訳文を生成し、各々を自立語化し情報量を比較している。

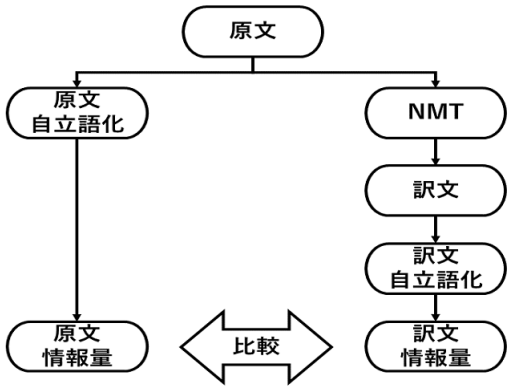

図 1: 提案手法の概要図

3.1 文の情報量

文を構成する単語の中で情報を持つと考えられる自立語を元に算出する。自立語 w が原文言語においてどの程度情報を持っているかという度合いを自立語 w の原文言語における存在確率から設定する。文としての情報量 S は自立語 w を組み合わせたものとして以下の式で定める。

$$S = \sum_{i=1}^{K} \log_2 \left(\frac{N}{w_{ni}} \right) \tag{1}$$

ここで N は原文言語の総自立語数である。w_n は原文言語中のそれぞれの自立語の出現回数であり、i は S 文中に現れる K 個の自立語の出現順である。

3.2 訳抜けの指標

式 1 の値は単語数が多い文や稀な単語が使われるほど高くなる傾向がある。単に原文の情報量 O と訳文の情報量 T の差をとった場合、訳抜けが少ない文であっても長文であることを理由に、割合として情報量の差が大きく出てしまうことがある。

そのため、訳抜けの指標 L としては $L = \frac{O-T}{O}$ として原文の情報量を元にした情報損失率を考える。実験では 50 %以上の情報が失われたものを訳抜けありとしている。

4 実験

4.1 設定

機械翻訳は英語の原文から日本語への訳文を対象とする。機械翻訳には Google の NMT(以下「GNMT」という) を用いた。ニューラルネットワークの学習の性質上、または Google の NMT モデル運用上の理由からか翻訳結果が一様に得られない場合がある。本研究では、2019 年 1 月 4 日における GNMT の結果を元にしている。

実験に用いる原文の内容は、英語のニュース文 3,000 文を対象とする。文量による効果をみるために、500 文ごとに提案手法を用いた実験も行った。図 2 は 1 文あたりの規模を日本語の自立語数で表し、その分布を示したものである。平均 16.3, 中央値 16 の自立語を含む文を対象にしている。

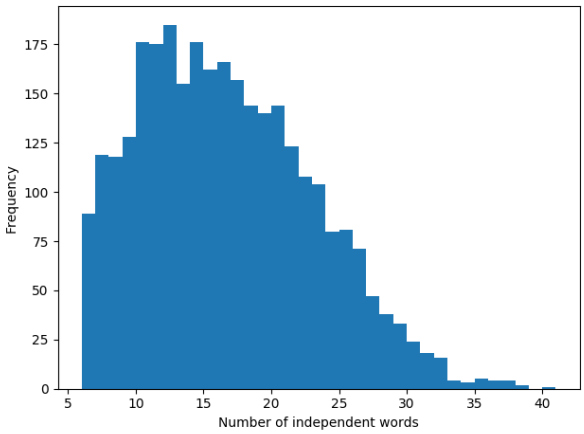

図 2: 訳文における自立語数分布

情報として扱う自立語について、訳文となる日本語は自立語の概念があるため形態素解析した品詞の助詞、助動詞を抜いたものを対象とする。原文である英語は、日本語の自立語に対応するものを Part of Speech の中から選んで用いている。

表 1 は、これらの設定で 3,000 文の自立語がどの程度になるかを示したものである。

表 1: 3,000 文の自立語数

	総数	種類数
英語原文	59,145	11,384
機械訳文	48,774	9,954

評価には真陽性率を用いる。ここで陽性とは訳抜けありのことである。提案手法は 50 %以上の情報が失われたものを陽性としている。提案手法によって検出された陽性を分母に、その中から真に陽性なものを分子にしたものが真陽性率となる。

真に陽性なものの判断は、原文と訳文を比較し人手で行う。真に陽性なものの基準は、原文から得られる情報が抜けているか否かを基準としている。誤訳は原文の内容を訳出しているので訳抜けとしていない。

4.2 実験結果

提案手法によってニュース 3,000 文の訳抜けを検出した結果、50 %以上の情報が失われた訳抜けありと見られるものは全体の約 2 %にあたる 54 文あった。真陽性（真陽性率）は、51 文 (94.4%) である。

表 2 は情報が失われたまたは増加したと判断された文の例である。

原文 1 は主要内容である発言部が全て訳抜けている。原文 2 は訳抜けの辻褄合わせに、撃たれた人物を撃った人物とする誤訳が起きている。どちらも網掛け部分の英文が訳されておらず、明らかな訳抜けを検出していることがわかる。原文 3 は本手法によって、情報が増加したケースである。「billions」という 1 自立語に対し、形態素解析で「何」、「十」、「億」、「ドル」と分けてしまっている。

文量による効果をみるための 500 文ごとに提案手法を用いた実験結果は図 3 である。それぞれ 3,000 文からランダムにサンプリングした文群を提案手法で評価し、100 回行った結果の平均値である。

真陽性率は各文量において、高い値をとっている。

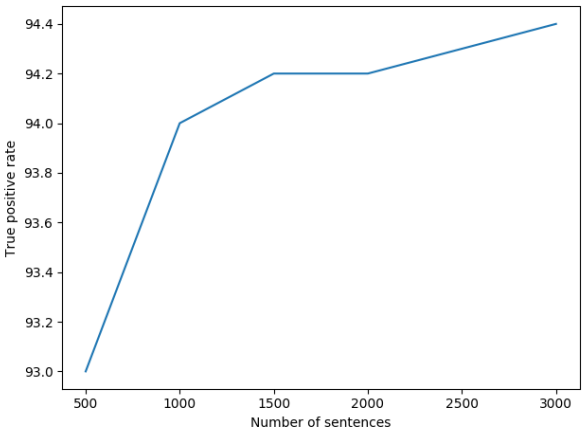

図 3: 提案手法による真陽性率

5 考察

英文 3,000 文を対象とした訳抜け検出において、提案手法は低コスト高精度な結果を得た。提案手法によって偽陽性となったものは陽性 54 文中に 3 文あった。主な要因は、原文が短文であることである。偽陽性の 3 文はそれぞれ自立語数が 7,10,11 の文であった。NMT は修飾の少ない短文における翻訳精度が高い。これは短文におけるエントロピーは低く、訓練された領域から外れにくいためであると推測される。また、提案手法は自立語の選定に起因して短文における情報損失精度が粗くなることがある。これは形態素解析によって得られる日本語の自立語概念と、英語の自立語らしい語の選出に言語間レベルで差異があるためである。他には、New York などの複単語表現 (Multi Word Expressions;MWEs) ひいては固有表現抽出 (Named Entity Recognition;NER) の問題が挙げられる。あるエンティティを 2 言語間で 1 対 1 に対応付けるとき、高い形態素解析精度が必要であり単語レベルを基本とする今回の手法ではこの問題への柔軟性が低い。これらの要因から短文における精度不足が生まれると考えられる。

現在は訳抜けの検出閾値を情報損失 50 %以上としているため検出の質は高い。しかし、検出の量として改善の余地がある。閾値 50 %では全体の 2 %弱を検出するに過ぎない。このため原文数を 100 程度で提案手法を試すとランダムサンプリングの結果によっては陽性検出 0、真陽性率を算出し得ないケースがある。

表 2: 情報量の差が顕著な例

	情報量	文
原文 1	280	”We should all do everything we can to see that, in Jo’s memory, we bring an end to the acceptance of loneliness for good,” Prime Minister Theresa May said in a statement.
機械翻訳	57	テレサ・メイ首相は声明で次のように述べている。
原文 2	403	The Russian ambassador to Turkey was shot in the back and killed at an Ankara art gallery Dec. 19 by an off-duty police officer who shouted “ Don ’t forget Aleppo ” and “ Allahu Akbar ” as he opened fire.
機械翻訳	177	トルコへのロシア大使は 12 月 19 日、アンカラ美術館で「アレッポを忘れないで」と「アッラーフ・アクバル」を発砲して怒鳴ったため殺害された。
原文 3	100	Beijing has poured billions into such ambitious scientific projects.
機械翻訳	151	北京はこのような野心的な科学プロジェクトに何十億ドルもの資金を注いできました。

6 おわりに

本論文ではニューラル機械翻訳の訳抜け検出に対して、原文の情報量ベースの手法を提案した。原文の単語が持つ情報量を元に文全体のあるべき情報量を仮定し、訳文の情報量と比較することで訳抜けを検出している。実験では計算コスト低く、明らかな訳抜けを検出している。

提案手法の改良としてまず挙げられる点は MWEs と NER である。現在は 1 エンティティの内容を日本語 1、英語 5 として自立語数をカウントするケースがある。MWEs を 1 エンティティとして認識する枠組みとして ngram などを採用し、NER によって 1 エンティティ性を高めることで、提案手法の精度向上が期待できる。同時に提案手法の検出閾値を下げて実験を行うために、人手による訳抜け判定作業を進め閾値が低い場合の知見も取り入れたい。

また、本研究は計算コストを軽い制約としているためニューラルネットワークを用いていないが、情報を持つ単語や品詞の選定についてはニューラルネットワークによる事前調査傾向を用いて重みなどを新たに設定することで手法の改良を試みたい。これと形態素解析のより詳細な結果を併用する方向で研究を進めたい。

近年、機械翻訳は人が受け入れやすい自然な文をもたらしている。しかし、実験結果表 2 の原文 2 に顕著であるが、自然な文とするための流れを重視するために事実の主客を入れ替えてしまうケースもある。人にとっての自然さを見せ始めたニューラルネットワークを人が理解し続ける必要があることを改めて示唆している。

参考文献

[1] Isao Goto and Hideki Tanaka. Detecting untranslated content for neural machine translation. In Proceedings of the 1st Workshop on Neural Machine Translation (ACL-2017), pp. 47-55, 2017.

[2] Ryuichiro Kimura, Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto. Effect on reducing untranslated content by neural machine translation with a large vocabulary of technical terms. In Proceedings of the 7th Workshop on Patent and Scientific Literature Translation, pp. 13-24, 2017.

[3] 中澤敏明. 機械翻訳の新しいパラダイム：ニューラル機械翻訳の原理. 情報管理, Vol.60, No.5, pp. 299-306, 2017.