課題番号	Q21D-01
課題名（和文）	半教師ありデータ拡張を用いたオープンソース英日ニューラル機械翻訳モデルの開発
課題名（英文）	Development of an Open-Source Japanese-English Neural Machine Translation Model with Semi-Supervised Learning
研究代表者	所属（学部、学科・学系・系列、職位） 先端科学技術研究科、情報通信メディア工学専攻　博士課程（後期） 氏名 ルスリ　アンドレ
共同研究者	所属（学部、学科・学系・系列、職位） 先端科学技術研究科、情報通信メディア工学専攻　教授 氏名　宍戸　真
	所属（学部、学科・学系・系列、職位） 氏名
	所属（学部、学科・学系・系列、職位） 氏名
	所属（学部、学科・学系・系列、職位） 氏名

研究成果の概要（和文）

本研究では、主に 2 つの研究成果を提案している。一つ目は、半教師の環境でバックトランスレーションと Transformers モデルに基づく multi-head attention モデルを用いて学習した複数の日英ニューラル翻訳モデル（NMT）の翻訳性能に関する実験報告である。まず、一般に公開されている 2 つの日英並列コーパス（JESC と JParaCrawl）を用いてモデルを学習し、それに加え、Business Scene Dialogue（BSD）コーパスから会話文を含む並列データセットを用いて fine-tuning を行った。BSD コーパスを用いた学習では、オリジナルのデータに加え、その逆翻訳データも使用した。二つ目は、日本語の翻訳において特に困難とされる anaphoric zero-pronoun を含む翻訳性能を評価するための specialized dataset の構築を支援するツールを開発した。その結果、いくつかのモデルは BLEU スコアで良好な結果を示したものの、anaphoric zero-pronoun を解決することは依然として困難なタスクであることが証明された。特に日本語の文章を翻訳する際に、NMT モデルの性能を評価する際に、specialized dataset の重要性を示唆している。さらに、アノテーションされた zero-pronoun データセットを用い、XLM-R という事前学習した多言語モデルを fine-tuning し、parallel dataset に anaphoric zero-pronoun 現象が含まれているかどうかを認識する分類モデルを学習した。このモデルは、本研究で開発した支援ツールを用いてアノテーションを行う際に、人間のアノテーターを支援するのに有用であると考えられる。

研究成果の概要（英文）

In this research, we provide two main contributions. The first contribution is an experimental report of the translation performance of several Japanese-English neural translation models (NMT) trained in a semi-supervised setting using back-translation and a multi-head attention model based on the Transformers model. The model is initially trained using two publicly available Japanese-English parallel corpora (JESC and JParaCrawl), then fine-tuned using a parallel dataset containing conversational sentences from the Business Scene Dialogue (BSD) corpus. When training using the BSD corpus, we used the original dataset, plus its back-translated version. Our second contribution is that we developed a support tool for building a specialized set to evaluate translation performance containing anaphoric-zero pronoun which is especially challenging in translating sentences written in Japanese. The results show that even though some models seem to perform well in BLEU score on a parallel dataset, resolving anaphoric zero-pronoun is still a difficult task. This implies the importance to have a specialized set when evaluating the performance of NMT models, especially in translating Japanese sentences. Additionally, using the annotated zero-pronoun dataset and fine-tuning the XLM-R pre-trained multilingual model, we trained a classification model to recognize whether a parallel sentence contains the anaphoric zero-pronoun phenomenon or not. This model can be helpful to assist human annotators when annotating using the support tool developed in our research.

1. 研究開始当初の背景

深層機械学習分野の進歩のおかげで、 自然言語を処理するコンピュー ターの能力は、 近年大幅に向上している。その中でも急速に成長している分野の 1 つは、 ニューラル機械翻訳(NMT)である。しかし、 深層機械学習を活用して NMT モデルを開発するには、多くのラベル付きデータが必要だ。手動でラベル付けされた多くのデータを収集する労力を最小限 に抑えるために、半教師ありデータ拡張を活用することが考えられる。データ拡張方法の１つは､逆翻訳という手法である。基本的には、 ターゲット言語のコーパスを別の言語に翻訳してから、それをまた元の言語に翻訳して、意味は似ているが内容が少し異なるコーパスを取得することで機能できる。ただし、これを使うことによって日英の NMT モデルはどれだけの精度で会話を通訳ことができるかを明確するには実験が必要である。なぜなら、日本語では特殊な言語現象が様々でそのうちの一つは anaphoric zero-pronoun という現象である。最後に、モデルやツールに開発することによって日英の L2 言語学習を支援するための翻訳モデルの有用性の見通しが得られると考慮される。

2. 研究の目的

まず、日本語を処理するための NLP ツールを研究する。さらに、最適の日英 NMT モデルを学習するための方法を調べ、精度を評価し比較する。オープンアクセスの parallel dataset や Transformers の結構や SentencePiece やバックトランスレーションなどという日英 NMT モデルを学習するための手法やリソースを考慮し、モデルの生成と実験を行う。それらに加え、日英モデルを構築するために、日本語特殊の現象を研究し、解決方法を探す。

3. 研究の方法

(1)日本語を処理するための解析ツールやトークナイザーを研究
(2)半教師データ拡張手法を考慮し、実装
(3)日英 NMT モデルを学習し、精度の評価と比較
(4)Anaphoric zero-pronoun の評価データセットを構築するための支援ツールを開発し、データアノテーションを行う
(5)構築した Zero-pronoun データセットを使って NMT の評価に zero-pronoun 現象の重要性を考慮

4. 研究成果

大きく分けて、本研究の成果は２つある。一つ目は、半教師の環境でバックトランスレーションと Transformers モデルに基づく multi-head attention モデルを用いて学習した複数の日英ニューラル翻訳モデル（NMT）の翻訳性能に関する実験報告である。まず、一般に公開されている 2 つの日英並列コーパス（JESC と JParaCrawl）を用いてモデルを学習し、それに加え、Business Scene Dialogue（BSD）コーパスから会話文を含む並列データセットを用いて fine-tuning を行った。BSD コーパスを用いた学習では、オリジナルのデータに加え、その逆翻訳データも使用した。二つ目は、日本語の翻訳において特に困難とされる anaphoric zero-pronoun を含む翻訳性能を評価するための specialized dataset の構築を支援するツールを開発した。その結果、いくつかのモデルは BLEU スコアで良好な結果を示したものの、anaphoric zero-pronoun を解決することは依然として困難なタスクであることが証明された。特に日本語の文章を翻訳する際に、NMT モデルの性能を評価する際に、specialized dataset の重要性を示唆している。さらに、アノテーションされた zero-pronoun データセットを用い、XLM-R という事前学習した多言語モデルを fine-tuning し、parallel dataset に anaphoric zero-pronoun 現象が含まれているかどうかを認識する分類モデルを学習した。このモデルは、本研究で開発した支援ツールを用いてアノテーションを行う際に、人間のアノテーターを支援するのに有用であると

考えられる。

5. 主な発表論文等

〔雑誌論文〕（計　2 件）

① Rusli, A. & Shishido, M. “Zero-Pronoun Annotation Support Tool for the Evaluation of Machine Translation on Conversational Texts”, 会誌「自然言語処理」Vol. 29 no. 2 (June 2022) オンライン
– *accepted, to be published*

② Rusli, A. & Shishido, M. “Utilizing Natural Language Processing to Develop an Interactive Web Platform for Practicing Text-based Conversational English as a Foreign Language”, The 10th European Conference on Language Learning (ECLL2022), September 2022. オンライン
– *accepted, to be published*

〔学会発表〕（計　2 件）

① Rusli, A. & Shishido, M. “On the Applicability of Zero-Shot Cross-Lingual Transfer Learning for Sentiment Classification in Distant Language Pairs”, 言語処理学会第 28 回年次大会（NLP2022), March 2022,オンライン

② Rusli, A. & Shishido, M. “Utilizing Natural Language Processing to Develop an Interactive Web Platform for Practicing Text-based Conversational English as a Foreign Language”, The 10th European Conference on Language Learning (ECLL2022), July 2022,オンライン
– *accepted, to be presented*