

論文 / 著書情報
Article / Book Information

題目(和文)	学習時と推論時における入力データの特徴の違いを考慮したニューラル機械翻訳モデルの学習手法
Title(English)	
著者(和文)	美野秀弥
Author(English)	Hideya Mino
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11827号, 授与年月日:2022年3月26日, 学位の種別:課程博士, 審査員:徳永 健伸,岡崎 直観,村田 剛志,宮崎 純,下坂 正倫
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11827号, Conferred date:2022/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

論文要旨
THESIS SUMMARY

系・コース： Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位 (専攻分野)： 博士 (工学) Academic Degree Requested Doctor of
学生氏名： Student's Name	美野 秀弥		指導教員 (主)： Academic Supervisor(main) 徳永 健伸 教授
			指導教員 (副)： Academic Supervisor(sub)

要旨 (和文 2000 字程度)
Thesis Summary (approx.2000 Japanese Characters)

機械翻訳は、ある言語で書かれた文（原言語文）を入力としてこれを異なる言語の文（目的言語文）に変換する技術であり、異なる言語を扱う国や地域の人々とのコミュニケーションや異なる言語で書かれた情報の理解などの効率化のための重要な技術として注目されている。近年は、深層学習を用いたニューラル機械翻訳により旧来の統計的機械翻訳と比較して翻訳精度が飛躍的に向上し、盛んに研究が行われている。ニューラル機械翻訳は事前に用意したデータで学習した翻訳モデルを用いて翻訳する。しかし、学習時と推論時とでデータの特徴に違いが生じることがあり、このために翻訳精度が低下する問題がある。本論文では、この問題に着目し、学習時と推論時（以下、翻訳時とする）で用いるデータの特徴が違う2つのニューラル機械翻訳のタスクを扱う。

1つ目は、ドメインタグを用いたニューラル機械翻訳である。翻訳モデルは翻訳対象のデータと同じ特徴を有する対訳データで学習することが理想である。しかし、対訳データの構築にはコストがかかるため、翻訳対象のデータとは特徴が違う対訳データを学習データとして利用することがある。特徴が違う対訳データを含めて学習すると、翻訳精度は低下することが知られている。この問題を解決するための手法の1つがドメインタグを用いてコーパスを区別して学習する手法である。本論文では、従来のドメインタグ手法を改善するため、翻訳対象が有する特徴、翻訳結果が有する特徴、翻訳対象と翻訳結果との関係に関する特徴、の3つの観点で特徴が違う対訳データを明示的に区別して学習する手法を提案した。評価実験の結果、ドメインタグを用いる従来手法と比較して提案手法の翻訳精度が向上することを確認した。

2つ目は、文脈情報を用いたニューラル機械翻訳である。ニューラル機械翻訳は、翻訳精度を上げるために、入力に翻訳対象の周辺の文である文脈情報を活用する手法がある。しかし、主に翻訳精度の向上に寄与する文脈情報は原言語側の周辺の文であり、目的言語側の周辺の文を文脈情報として用いる従来手法は翻訳精度の低下が報告されていた。本論文では、目的言語側の文脈情報を用いる従来手法の問題が翻訳モデルの学習時と翻訳時で用いる文脈情報のデータの特徴の違いにあると仮定し、学習時と翻訳時で用いる文脈情報の特徴の違いによる影響を緩和するためのデータ制御方法を提案した。評価実験の結果、目的言語側の前文を利用する従来手法と比較して、提案手法の翻訳精度が向上することを確認した。

本論文では、上記のように、ドメインタグを用いたニューラル機械翻訳と文脈情報を用いた2つのニューラル機械翻訳のタスクにおいて、学習時と翻訳時で入力データの特徴が違うために翻訳精度が低下していることを示し、各タスクに対して翻訳精度が向上する有効な手法を提案し、評

価実験により提案手法の有効性を示した.

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

(博士課程)
Doctoral Program

論文要旨

THESIS SUMMARY

系・コース : Department of, Graduate major in	情報工学 知能情報	系 コース	申請学位 (専攻分野) : Academic Degree Requested	博士 Doctor of	(工学)
学生氏名 : Student's Name	美野 秀弥		指導教員 (主) : Academic Supervisor(main)	徳永 健伸 教授	
			指導教員 (副) : Academic Supervisor(sub)		

要旨 (英文 300 語程度)

Thesis Summary (approx.300 English Words)

In recent years, neural machine translation (NMT) using deep learning has improved translation quality compared to statistical machine translation. NMT uses a translation model trained on data prepared in advance. However, there is a problem that the features of data which differ between training and inference, resulting in a degradation of translation quality. We focus on this problem and deal with the following two NMT tasks.

The first is NMT using domain tags. Basically, the translation model should be trained on bilingual data that has the characteristics of data to translate. However, since the construction of bilingual data is costly, bilingual data with different features from the ideal data is often used. It causes the degradation of translation quality. In this paper, we proposed a training method that explicitly distinguishes bilingual data with different features in three aspects: features of the translation target, features of the translation result, and features related to the relationship between the translation target and the translation result. Experimental results show that the proposed method improves the translation accuracy compared to the conventional method using domain tags.

The second is NMT using contextual information which is the surrounding sentences of the translation target. Although the context information of the surrounding sentences of the source language contributes to the improvement of translation quality, the conventional methods that use the surrounding sentences of the target language have been reported to degrade translation quality. In this paper, we propose a data control method to mitigate the effect of the difference in the features of the context information used for training and translation of the translation model, assuming that the problem of the conventional method using the context information of the target language side is the difference in the features of the data of the context information used for training and translation. As a result of evaluation experiments, we confirmed that the proposed method improves translation quality compared to the conventional method.

備考 : 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意 : 論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。

Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).