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)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		美野 秀弥	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	徳永 健伸	教授	審査員	下坂 正倫	准教授
	審査員	岡崎 直観	教授			
		村田 剛志	教授			
		宮崎 純	教授			

論文審査の要旨（2000字程度）

本論文は「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」と題し，和文6章から構成されており，ニューラル機械翻訳においてモデルの学習に使うデータと翻訳対象の性質が異なる場合でも，高い翻訳性能を維持する手法を提案している．

第1章「序論」では，本論文の背景と目的について述べている．ニューラル機械翻訳で質の高い翻訳結果を得るためには，大量の学習データを必要とするが，翻訳対象とするドメインの学習データを常に大量に用意できるとは限らず，翻訳対象とは異なるドメインのデータを学習時に併用することがある．しかし，異種データの併用により翻訳性能が低下することが知られている．このような背景をふまえ，異種データの利用による翻訳性能の劣化を抑制しつつ，学習データを増やすことによる翻訳性能の向上を実現することがニューラル機械翻訳で重要であると述べ，本研究の目的がこの相反する目標を同時に実現する手法を提案することであると述べている．そのために，まず，学習時と翻訳時のデータの特徴のどのような要因に着目するかを整理し，(a) 翻訳対象，(b) 翻訳結果，(c) 翻訳対象と翻訳結果の関係，(d)外部情報の4つの要因にまとめている．これらの要因を網羅するように2つの手法，ドメイン・タグを利用した要因(a)，(b)，(c)の違いの解消および要因(d)の違いを低減する文脈を考慮したニューラル機械翻訳を提案すると述べている．

第2章「ニューラル機械翻訳」では，現在，機械翻訳の主流となっているニューラル機械翻訳の基本原理および翻訳結果の評価方法について説明している．

第3章「関連研究」では，翻訳モデル学習時に学習データのドメイン・タグを利用する過去のニューラル機械翻訳，および原言語側，目的言語側の先行文脈を利用するニューラル機械翻訳の先行研究について述べ，それらの研究の問題点を指摘している．

第4章「複数のドメイン・タグを用いたニューラル機械翻訳」では，異種データを利用して学習をおこなう際に，先行研究で提案されているコーパスタグを使う手法，ノイズタグを使う手法，両方を併用する手法について概観し，いずれの手法も学習時のデータと翻訳対象の性質の違いを表現するには不十分であると述べている．それをふまえ，本論文では原言語側のデータの特徴，目的言語側のデータの特徴，さらに原言語と目的言語の関係に着目したタグを付与することにより，データの特徴の違いに関するより粒度の細かい情報を表現できると主張している．提案手法の有効性を確認するために，まず，ドメイン・タグを付与せずに異種データを追加して学習データ量を増やしても翻訳性能が向上しないことを確認し，ドメイン・タグの有効性を示している．次に，従来手法と提案手法によるドメイン・タグを付与して学習をおこなった結果，翻訳性能が改善されることを確認している．特に提案手法が従来手法よりも改善の幅が大きいことを示している．提案手法では，ドメイン・タグの種類が多いので，データの過疎性が問題となる可能性がある．そこで，同一ドメインの学習データ量を変化させることにより，提案手法の有効性は同一ドメインの学習データ量に影響を受けること，また，ドメイン・タグの種類を増やすだけではなく，タグを適切に組み合わせることが性能向上に必須であることを明らかにしている．

第5章「目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳」では，代名詞などの照応の解消やスタイルの選択など，文脈を考慮しないと適切な訳が得られない場合に対処する文脈を用いたニューラル機械翻訳について，まず，先行研究で提案されている従来法を概観し，その問題点について述べている．従来，文脈情報として原言語側の先行文脈を使うと翻訳性能が向上するが，目的言語側の先行文脈を使うと翻訳性能が低下するとされていた．従来手法では，学習時には目的言語の参照訳を使って学習し，翻訳時には機械翻訳結果を使って推論をおこなっている．本論文では，両者のデータの性質の違いが翻訳性能の劣化を招いているという仮説のもと，学習時にも目的言語の先行文脈の機械翻訳結果を使う手法と目的言語の参照訳と機械翻訳結果をカリキュラム学習

の考え方を取り入れて段階的に併用する手法を提案し，いずれの手法も原言語側の先行文脈を使う従来手法よりも上回る性能を達成できたと述べている．特にカリキュラム学習を取り入れた手法の改善幅が大きく，人手による事例分析の結果でも照応を含む文を正しく翻訳できた事例数が 49 から 63 に増えたと報告している．

第 6 章「結論」では，ニューラル機械翻訳における学習データと翻訳対象の性質の違いに対処しながら学習データを増やすことが，翻訳性能を向上させる上で重要であることを指摘し，そのために注目すべきデータの違いと特徴付ける重要な要因として，翻訳対象，翻訳結果，両者の関係，外部情報の 4 つをあげている．これらの違いを網羅したニューラル機械翻訳としてドメイン・タグを利用する手法と目的言語側の先行文脈として参照訳と機械翻訳結果を段階的に併用するカリキュラム学習を取り入れた手法を提案し，評価実験を通して提案手法の有効性を示したことが本論文の貢献であると述べている．

以上要するに，本論文は学習時と翻訳時のデータの性質の違いを考慮したニューラル機械翻訳の手法を提案し，その有効性を評価実験により示している．本論文の成果はニューラル機械翻訳の翻訳性能改善のための基盤技術として位置付けることができ，工業上貢献するところが大きい．よって，本論文は博士(工学)の学位論文として十分価値あるものと認める．