別紙様式第１５号（論文内容の要旨及び論文審査の結果の要旨）

氏名（本籍）	ZHANG JINYI（中華人民共和国）
学位の種類	博　士（工学）
学位授与番号	甲第５７７号
学位授与日付	令和２年３月２５日
専攻	電子情報システム工学専攻
学位論文題目	日中・中日ニューラル機械翻訳のための文字特徴情報の利用とコーパス拡張手法 （Additional input character features and corpus augmentation for neural machine translation between Japanese and Chinese）
学位論文審査委員	（主 査）教授　草刈　圭一朗 （副 査）教授　山口　忠　　准教授　松本　忠博

論文内容の要旨

本論文は，文字レベルの日中・中日ニューラル機械翻訳の精度向上を目的として，入力文中の各文字の特徴情報を取得して利用する手法，および，学習データである対訳コーパスを増やすデータ拡張手法について研究し，まとめたものである。

近年，ニューラル機械翻訳の登場により，従来主流であった統計的機械翻訳に比べ，翻訳精度や訳文の流暢さが格段に向上した。しかし，質の高い翻訳結果を得るには，統計的機械翻訳以上に訓練データとなる対訳コーパスが大量に必要であり，学習や翻訳にもより多くの時間を要する。通常の単語レベルのニューラル機械翻訳では，訓練時間が膨大になるのを防ぐため，低頻度語を一律に UNK（未知語を表す特殊な語）に置き換えるか，単語を部分文字列に分割するなどして語彙サイズを抑える必要がある。しかし，標語文字（logogram）である漢字を共有する日本語と中国語では文字レベルの訓練・翻訳が有効であり，これにより語彙サイズの問題や単語分割の誤りなども回避できることから，本研究では文字レベルの日中・中日ニューラル機械翻訳を研究対象とした。

単語レベルのニューラル機械翻訳では，単語の特徴ベクトルに，品詞などの言語学的情報を特徴情報として追加することで翻訳精度が向上することが知られている。本研究では，文字レベルのニューラル機械翻訳においても文字に関する何らかの情報を付与することで翻訳精度が改善される可能性があると考え，漢字の部首を文字特徴情報として追加する手法を提案した（第４章）。部首は，字書において文字の分類や検索に用いられる漢字の構成要素である。漢字はその成り立ちから，象形文字，指示文字，会意文字，形声文字に分類されるが，その多くを占める形声文字において，物事の類型や意味範疇を表す構成要素（意符）が部首となっている場合が多い。そこで本研究では，部首が持つ意味的な情報が翻訳精度の向上につながると考え，入力特徴情報に部首を加えた。漢字以外の文字にも便宜的に部首を定めた。手法の評価のため，ASPEC（アジア学術論文抜粋コーパス）に含まれる日中対訳コーパスを用いた翻訳実験を行い，翻訳精度が改善されることを確認した。

また，ニューラル機械翻訳には大量の訓練データが必要だが，日英対訳コーパスなどに比べると日中対訳コーパスの量は少なく，十分とは言えない状況にある。本論文では，対訳コーパスの不足を補う方法として，対訳文対の分割と逆翻訳によるコーパス拡張手法を提案している（第５章）。対訳文対の分割には，単語アラインメント情報（単語間の対応関係）を利用し，その補正のために，日中の漢字の対応表を利用した。目的言語側の部分文を逆翻訳して，擬似原言語部分文を生成し，元の原言語文の一部をこの擬似原言語部分文と置き換えることで，原言語文の複数のバリエーションを取得する。これらの擬似原言語文と元の目的言語文を対にすることで，１つの対訳文対から複数の擬似対訳文対が得られる。上述の日中対訳コーパスからに抽出した対訳データに提案手法を適用して得られた拡張コーパスを用いて評価実験を行った結果，提案手法の有効性が示された。

論文審査結果の要旨

ニューラル機械翻訳によって生成される訳文の品質は，学習データとなる対訳コーパスの量や質，内容に強く依存している。日中対訳コーパスは，日英対訳コーパスなどと比較するとその量はまだ少

なく，領域も限られる。本論文では，日中・中日ニューラル機械翻訳の品質を向上させることを目的として，（1）入力文中の各文字（主に漢字）の類型・意味範疇に関わる情報を文字の追加特徴情報として入力に加える手法，及び，（2）対訳文対から対訳部分文対への分割とその逆翻訳により既存の対訳コーパスを拡張する手法を提案している。対訳文対の分割は，既存のツールを使用して得られた単語アラインメント情報（原言語文と目的言語文の間の単語の対応関係）をもとに行うが，本論文では日本語と中国語の漢字の対応表を用いてアラインメント情報を補正する手法を導入している。提案されたコーパス拡張手法は既存の対訳文をもとに複数の擬似対訳文を生成するため，他のコーパス拡張手法との併用も可能となっている。いずれの手法に対しても，アジア学術論文抜粋コーパスに含まれる日中対訳データセットを用いて翻訳実験を行っており，翻訳品質の一般的な評価指標である BLEU（コーパス拡張手法については，BLEU と TER）を用いた翻訳結果の評価によりそれぞれの手法の有効性が確認されている。

審査委員会において学位申請論文を慎重に審査した結果，本論文は工学的な価値が高く，博士（工学）の学位論文として十分な内容を持つものと判定した。

最終試験結果の要旨

学位審査委員会は，本論文の主要部分が査読付きジャーナル及び査読付き国際会議プロシーディングに掲載された 2 編の論文を基に構成されており，学位論文の基礎となる学術論文に関する判定基準「学術誌に 2 編以上．ただし，このうち 1 編は国際会議のプロシーディングや国際会議に関連して発行された査読付き論文集であってもよい」を満たしていることを確認した。また，令和 2 年 1 月 27 日に開催された学位論文公聴会における発表内容及び質疑に対する回答状況などを踏まえて審査を行い，最終試験の結果を合格と判断した。

発表論文（論文名、著者、掲載誌名、巻号、ページ）

1. Corpus augmentation for neural machine translation with Chinese-Japanese parallel corpora, Jinyi Zhang, Tadahiro Matsumoto, Applied Sciences, vol.9, no.10, 2036, pp.1-16, 2019, doi:10.3390/app9102036.
2. Improving character level Japanese-Chinese neural machine translation with radicals as an additional input feature, Jinyi Zhang, Tadahiro Matsumoto, Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017, IEEE Part Number: CFP1744I-USB, ISBN: 978-1-5386-1980-3, pp.172-175, 2017.