氏名（本籍地）	阿部 香央莉（あべ かおり）　（青森県）
学位の種類	博　士（情報科学）
学位記番号	情　博　第　７９７　号
学位授与年月日	令和　５年　３月２４日
学位授与の要件	学位規則第４条第１項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）システム情報科学専攻
学位論文題目	Application-oriented Machine Translation: Design and Evaluation （実応用を志向した機械翻訳システムの設計と評価）
論文審査委員	（主査）教　授　乾 健太郎 教　授　張山 昌論　　教　授　伊藤 彰則 教　授　鈴木 潤

論 文 内 容 の 要 旨

第１章　Introduction

近年の機械翻訳の発展は目覚ましく、それに伴って実世界における応用シーンでの利用が拡大している。Web ブラウザ上での翻訳機能や Google 翻訳・DeepL 翻訳などの機械翻訳アプリを用いることで、例えば、外国語で記された Web 上の記事などのコンテンツを読んだり、あるいは上記翻訳ツールの出力を補助として外国語で特許や科学論文などの特定の文書を執筆したりするなどの活用が盛んに行われている。また近年では、Zoom や Google Meet などのオンライン会議アプリや YouTube などの動画サイトなどにおいて、音声認識と機械翻訳技術を組み合わせた自動同時通訳の試みも浸透しつつある。

しかし、そのような機械翻訳の実世界利用が拡大している中で、現状の機械翻訳分野には大きく 2 つの課題がある。1 つは、汎用的な機械翻訳性能向上に伴い、実世界においてユーザ（個人・法人）個別の詳細な希望を反映することができるような機械翻訳システムの需要が伸びていること、もう 1 つはシステム出力内に依然として含まれる不適切な出力により、機械翻訳システムの利用でコミュニケーションの齟齬や誤解を生みかねないことである。これらは、今後も機械翻訳市場が益々拡大すると予見されている中で、極めて対処の重要性が高い課題である。本論文では、この 2 つの課題に対し、機械翻訳のシステムデザインおよび頑健な評価基盤の構築という 2 軸の観点からのアプローチを行う。

第２章　Background

本章では、本論文における 2 つのアプローチである機械翻訳のシステムデザインおよび頑健な評価基盤に関する背景知識を説明する。

システムデザインの背景知識として、まず「オプション機能」およびそれらを組み合わせる方法論の 2 つを導入している。機械翻訳分野における様々な問題（例：低資源言語や特定の用語の訳出など）に対して個別に提案されてきた、一般的な sequence-to-sequence モデルに対して入力情報やエンコーダ・デコーダの仕組みの変更を行うアーキテクチャを総称して「オプション機能」と本稿では呼称し、このオプション機能と見做せる研究例の紹介を行った。加えて、現実世界における課題の実例（例：新薬に関する特許の国際出願や異言語間の対話支援）に対して、それらのオプション機能を組み合わせたシステムデザインの例示を行った。第 3 章では、この実世界での課題の例のうち 1 つを掘り下げる形で、多方言翻訳システムのデザインを紹介する。

また、頑健な評価基盤に関する背景知識として、機械翻訳分野における従来の自動・人手評価基盤の枠組みを体系的に整理した。具体的には、最もメジャーな n-gram（表層一致）ベースの自動評価指標

BLEU の問題点を元に USE, SimCSE などの意味的類似度を考慮した指標が多数提案されてきたこと、また近年では、より厳密にシステム出力の良し悪しを評価するため、人手評価に対しても見直しの流れがあることを説明した。これは、第 4 章や第 5 章で行う現状の評価基盤への分析の理解を深める目的に基づいて記述したものである。

第 3 章　Design：Japanese Multi-Dialect Translation as Use Case

第 2 章で複数提示した、実応用の課題を解くための機械翻訳システムデザインの例の一つとして、日本語の多種多様な方言を標準語に翻訳する多方言機械翻訳システムの設計およびその実装を行なった。一つ一つの方言に関するデータが低資源な状況にある、かつ方言の種類が多様であるという課題を抱える日本語方言の処理において、方言間の共通性および方言と標準語間の共通性（例：語彙・用いる文字種・語順の共通性等）を利用し、多言語翻訳の仕組みおよびその効果を助長する 3 種類のオプション機能を導入して機械翻訳システムのデザインを行った。3 種類のオプションの内訳は、それぞれ (1) 特殊トークンを用いた多言語翻訳、(2) 文字（音韻）レベルのエンコーディング、および (3) 語順固定の翻訳方式となっている。特殊トークンを用いた多言語翻訳は、低資源な状況下ではあるものの多くの点で類似している方言の特徴を包括的にシステムに学習させることで、1 つのシステムで複数の方言の翻訳を実現する狙いがある。また、文字レベルのエンコーディングに関しては、方言間の違いが主に音韻の変化に起因することから、この違いをより捉えやすくするという狙いがある。また、語順固定の方式は、方言・標準語間で語順が変化しないことから、アーキテクチャ側でこれを事前に固定し、システムが学習中に見るインスタンスの量を増加させることで、低資源なデータを効率よく学習するという狙いがある。これら 3 つのオプション機能を組み合わせることで、48 種類もの方言に対処可能な機械翻訳システムを構築した。

実験では、導入した 3 つのオプション機能がそれぞれ性能向上に深く関連していることを ablation study によって示した。また、システム内部の各方言を表す分散表現の分布が方言学で提唱されている各方言の分布に似た傾向を示していることから、本章にて設計されたシステムが方言毎の特徴を捉えて翻訳を行っているということを示した。これらの実験や分析を通して、実世界の課題に応じて適切なオプション機能を選択したシステムデザインを行うことの重要性を示した。

第 4 章　Evaluation (1)：Comprehensive Analysis of Semantic Metrics for MT

機械翻訳システムの性能向上を促進するためには、高コストな人手評価の代わりとなる低コストな自動評価指標が欠かせない。しかし、従来からシステム評価に盛んに用いられている n-gram ベースの自動評価指標 BLEU は、表層的には異なるものの意味的には許容可能な訳出に対し低いスコアを付与してしまうという問題が指摘され続けている。そこで、近年では参照訳との意味的な類似性を考慮可能な自動評価が強く求められている。

しかし、近年提案されている意味的類似度を考慮した自動評価指標は、汎用的な意味的類似度を評価するベンチマークタスク Semantic Textual Similarity, 通称 STS 上での評価を基に提案された評価指標と、機械翻訳評価（MT Evaluation, MTE）などのより実応用シーンを前提としたタスク上での評価を基に提案された評価指標の大きく 2 種類があり、どちらがより適切に意味的類似度を評価できるかが明らかになっていない。また、実際に各指標を 2 種類のタスクで評価してみると、その性能は各タスクによって大きく異なることが判明した。本章では、近年提案された自動評価指標（およびベースラインとなる指標）に関して、それらの性能がタスク・データセット中のどのような要因に影響を受けるのかを突き止めるため、STS および MTE という 2 種類のタスクデータセットおよびドメインなどの各要因を調整したサブセット上での性能の差を基に体系的な分析を試みた。

23 種類の自動評価指標に対する 2 タスク上での性能分析の結果、ドメインの違いや類似度粒度の

違い、利用シーンの違い等の要因が現状の自動評価指標の予測結果に影響を与えることを示した。各評価指標がこれらの要因に影響を受けることは、実験時とは異なる様々なドメインやシーンが考えられる実世界での利用において十分に考慮すべき点である。また、自動評価指標がこれらの要因に影響されることを踏まえ、現状の指標の評価において用いられているベンチマークデータセットに対するデータセット分布やアノテーションスコアのバイアスについての懸念や問題点を指摘した。

第５章　Evaluation (2) : Terminology-Focused Evaluation

実応用での機械翻訳の活用先の一つに、特許や科学論文など外国語（主に英語）での執筆の補助が挙げられる。このような、専門用語や特定の固有名詞などの「用語」が頻出する例において、その用語をユーザの希望（制約）に則して適切に出力できる機械翻訳システムの需要が高まっている。その需要に応えるべく、ユーザから事前に用語の制約を与えられた状態で翻訳を行う機械翻訳モデルの枠組みは多数提案されているが、それを統一的に評価するための評価基盤が不足している現状がある。

そこで、本章ではこのようなモデルの評価基盤として、「Restricted Translation Task（制約付き翻訳タスク)」というシェアードタスクを開催し、そのタスクの設計およびデータセット・評価指標の構築を行なった。タスク設定は、事前に各文に対して制約語彙（Restricted Vocabulary)となる用語のフレーズが与えられており、機械翻訳システムは原文とその用語のフレーズを入力として受け取り、出力としてその制約語彙を含む対象言語側の翻訳文を出力として返す、という形式となっている。この時、制約語彙が出力に含まれていないまたは意味的には正しくても制約語彙として与えられていない語彙が出力に含まれていた際は、その出力を不正解とする。このタスクを実現するためのデータセットとして、科学論文ドメインの並列コーパス ASPEC を元に、主に人手による専門用語のアノテーションを行い、制約語彙を文ペアごとに付加した。また、簡易的な自動評価指標として、BLEU と制約語彙の exact match を組み合わせた指標を考案した。具体的には、システム出力のうち制約語彙が含まれていない例を exact match によってフィルタリングし、フィルタリング後の出力と参照訳との BLEU を計算するという方式をとった。この自動評価は、出力の制約語彙に関する適切さと翻訳としての質の両方を同時に評価したスコアを算出するという狙いがある。

この評価基盤をもとにシェアードタスクを 2 年間に渡って開催し、提出されたシステム出力および評価結果の分析を行った。結果として、考案した自動評価は現状のシステムランキングには満足な性能を発揮したが、用語の出力位置が誤っているなどの厳密なエラーを捉えきれていないということや、人手評価のスコアに専門用語が複数出現する場合に妥当でないと見られるものがあるということなどが判明し、今後の課題として用語を扱う機械翻訳タスクに特化した評価の枠組みを再検討すべきであるということがわかった。

第６章　Conclusions

本論文では、実応用シーンで活躍する機械翻訳システムを考慮した場合の 2 つの課題、ユーザ個別の希望に対応する機械翻訳の需要および誤解を導きかねない不適切な出力を踏まえ、「個別のシーンに対しカスタマイズ可能かつ誤解を生まない機械翻訳システム」の構築を目指した。また、それを実現するための 2 軸のアプローチ「機械翻訳のシステムデザイン」および「頑健な評価基盤」を整理し、その具体的な取り組みを紹介した。

今後の発展として、ユーザ自身がカスタマイズ可能なシステムデザインの方法論を探究することや、評価結果に影響を与える要因を考慮して評価基盤の改良を実践すること、また、機械翻訳システムによる誤解を防ぐ試みとしてユーザ自身が機械翻訳システムの出力のエラーを認知できるような方法論の提案などが考えられる。

論文審査結果の要旨

機械翻訳は、近年技術が大きく発展し、実世界での利用も拡大しているが、未解決の重要課題が少なくとも 2 つある。1 つは個別のユーザの希望を反映するためのカスタマイズ性の実現、もう 1 つはコミュニケーションの齟齬を生む致命的な翻訳誤りの回避である。本研究では、機械翻訳におけるこれら 2 つの課題に対し、システムデザインおよび頑健な評価基盤の構築という 2 つの観点からアプローチした。本論文は一連の成果をまとめたもので、全編 6 章からなる。

第 1 章は序論である。

第 2 章では、機械翻訳において提案されているアーキテクチャを選択可能なオプションとしてみなすシステムデザインの方法論について論じるとともに、翻訳結果を人手評価・自動評価する枠組みを体系的に整理し、本研究の背景知識を説明している。

第 3 章では、実応用を指向した機械翻訳のシステムデザインの一例として、多方言機械翻訳システムの設計およびその実装を報告している。多言語翻訳の仕組みを導入することにより、少量の訓練データで 48 種類もの方言を扱えるシステムを実現した。構築したシステムが方言の体系を暗黙的に学習しているといった興味深い分析も示されており、高く評価できる。

第 4 章では、機械翻訳の自動評価の問題を取り上げ、意味的類似度に基づく評価指標の評価性能がドメインや類似度の粒度、利用シーンの違いによる影響を受けることを複数のベンチマークデータセット上の組織的な実験によって解明し、既存の自動評価指標の問題点を明らかにした。一連の分析結果は当該分野の研究者にとって示唆に富むものであり、高く評価できる。

第 5 章では、専門用語等の「用語」が適切に翻訳できるかを評価するための基盤として、用語翻訳タスクを新たに設計し、データセットと評価指標を構築している。タスクの結果分析を通して翻訳における用語の扱いに重要な示唆を与えるもので、高く評価できる。

第 6 章は結論である。

以上、本論文は実応用を志向した機械翻訳システムにおける 2 つの課題に焦点を当て、その改善のための具体的なアプローチおよび今後の方向性を示したものであり、情報科学の発展に寄与するところが少なくない。よって、本論文は博士（情報科学）の学位論文として合格と認める。