早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論　文　題　目

A Study of Translation Equivalences at Various Levels of Granularity for Chinese-Japanese Technical Translation

申　請　者
Wei YANG

情報生産システム工学専攻
用例翻訳•言語処理研究

2017 年 07 月

日本と中国では毎年それぞれの言語で多量の技術文書や特許が公開されている。特許を例とすると、2013年に日本は特許付与件数で世界第一位だったが、2015年には中国が特許付与件数および特許出願件数（100万件超）でも世界第一位となった。（世界知的所有権機関（WIPO）最新統計データ）このような状況の下、両国で書かれた特許や技術文書を双方向で翻訳する必要性がある。そのため、日本では日本科学技術振興機構（JST）が、中国では国家知的所有権事務所（SIPO）が、それぞれ独立した助成で、日中・中日機械翻訳プロジェクトを開始した。日本語と中国語の両言語間の翻訳に適し、かつ技術文書や特許で出現する専門用語にも対応できるようにするためには、以下の３つの問題を解決する必要がある。第一の問題は、機械翻訳に不可欠な資源であるパラレルコーパスの不足である。パラレルコーパスは機械翻訳の訓練データであり、文のレベルで対訳関係を持つ文の対の集合である。本学位論文の第４章で本問題を扱う。第二の問題は、単語アライメントの非一貫性である。現在の機械翻訳アプロー　では、単語が処理単位となっているが、日本語と中国語は単語境界が明示されない言語であるため、前処理としてそれぞれの異なる分かち書きツールを適用して分かち書きを行うことにより、単語アライメントの非一貫性が生じ翻訳品質に悪影響を与えている。この一貫性問題は第２章で扱う。第三の問題は、専門用語の対訳ミスである。技術文書では高度な専門用語が多く出現し、単語の対訳関係を取るツールにミスが多く、これも翻訳品質へ悪影響を与えている。この専門用語の問題は第３章で扱う。本学位論文では上記の三つの問題を解決することを目指して、高精度で対応づける手法を確立するため、「字」と「単語」および「文」という３つの粒度で日本語と中国語との対訳関係を検討し研究を行っている。

以下に、本学位論文の構成と各章ごとの評価を述べる。

第１章[Introduction]では、研究の背景、従来技術について説明し問題提起をする。問題を解決するためにとったアプロー　を紹介し、機械翻訳分野への貢献について記述している。

第２章[Chinese and Japanese characters]では、字、ただし漢字の粒度での研究について述べている。本章で扱う問題はオープンアクセスでアクセスできる日中電子辞書の不足問題を扱っている。NICTがEDRという大規模（約30万見出し語を含む）辞書を人手で構築したが、この資源はライセンス付きのものである。本章では字の粒度で漢字変換テーブルを構築する方法を提案し、従来技術の辞書自動構築においてその貢献を測定した。日本語と中国語の間では、同じ意味を持つ文字が多いため、オープンアクセスのユニコード資源から提案した自動的な方法で得られた日本語・中国語漢字変換テーブルの妥当性は98.5%である。電子辞書を自動構築する際、パラレルコーパスが少ない場合、中間言語として第３ヶ国語を経由する手法を用いる。行なった予備実験では従来技術のone-time-inverse consultation手法では見出しの品質は 76%となり、漢字変換テーブルでフィルタリングをかけた結果、品質

を 82%まで向上させることができた。見出し語数は、4 万 1 千個以上となり、その新規性を評価するため、EDR 辞書と比較した結果、81%の見出し語が EDR 辞書に存在しなかったことを明らかにした。

本実験では自動的に得られた単語の対訳関係の品質も、新規性も高く、日本語中国語漢字変テーブルの効率を証明することができた。従って、第 3 章と第 4 章では、より広い粒度での対訳関係をフィルタリングするため、本章で構築した漢字変換テーブルと辞書を利用する。

第 3 章[Monolingual and bilingual term extraction for re-tokenization in SMT]では、単語、ただし専門用語の粒度の研究について述べている。特許や技術文書の翻訳では、専門用語の翻訳が非常に重要である。本章で扱う問題は日本語・中国語間のオープンアクセス電子専門用語辞書の不足問題である。この問題に対して、訓練データから自動的に専門用語の対訳関係を取ることを提案する。しかし、多くの専門用語は分かち書きによって複数語になり、また、それぞれの言語の分かち書きの結果にも差異がある。さらに、一般的に専門用語の出現頻度は低い。従来技術のアライメントツールでは、出現頻度の低い複数語表現の対訳関係を取ることが不十分であり、複数語専門用語対訳関係を取るのは困難である。そこで二段階での前処理を提案する。第一段階では、従来技術（C-value）で複数語専門用語を位置付けた上で、これを一単語にする（re-tokenisation）。第二段階では、re-tokenisation されたデータをアライメントするため、出現頻度の低い単語に優れているツールを利用する。その後、通常の翻訳実験を行う。最良条件を決めるために、対訳関係推定値の様々な閾値で実験を行い、また re-tokenisation を訓練データだけではなく、 ュニングデータにも行った。最良条件では BLEU の 1 ポイントの向上が確認できた（本報告書では、発表する翻訳品質向上は全て、信頼間隔を確認した上、p 値<0.01 で統計学的に有意である）。対訳関係を高めるため、漢字変換テーブルを利用した結果、BLEU の 1.5 ポイントの向上ができた。また、一単語と複数語専門用語の関係を取ることによって、BLEU の 2 ポイントのさらなる翻訳品質の向上が得られた。翻訳品質だけではなく、対訳関係の妥当性を検討した。その結果、複数語と複数語専門用語の関係だけで最大妥当性は 80%、漢字変換テーブルの使用で 93%、一単語・複数語関係を加えると 95%まで向上させることができた。要するに、本章では、分かち書きが専門用語の対訳関係推定に与える悪影響を re-tokenisation で減少させることができることを明らかにした。

第 4 章[Quasi-parallel data construction]では、文、ただし短い文の粒度での研究について述べている。本章で扱う問題は日本語・中国語間のパラレルコーパス不足問題である。パラレルコーパスにある文対の量とその対訳関係の品質が最終的に翻訳品質に大きな影響を与えることは明らかである。しかし、日本語・中国語間のパラレルコーパスは少ない状況にある。オープンアクセスのものは殆どない。しかしながら、日本語、中国語のオープンア

クセス単言語資源は決して少なくはない。本章では、単言語のデータに基づき、日中パラレルコーパス自動構築の新しい手法を提案する。ただし、パラレルではなく、クアジパラレルコーパスを構築する。単言語で、（日本語、中国語共に独立した)文の対を書き換えルールの例としてクラスタリングし、その書き換えルールで多量の新たな文を生成する。生成された新たな文の間の対訳関係を推定するため、書き換えルールの対訳関係推定に基づく手法を提案した。得られた対訳関係が高い文対の集合はクアジパラレルコーパスと呼ぶ。クアジ（擬似的）と呼ぶ理由は、完璧な翻訳関係を持っていないためである。また、生成される文には流暢性と文法性の問題があるため、篩をかける必要があり、様々な実験で、対訳関係の正しさより文法性が重要であることを明らかにした。ここは、提案した固定長文字列に基づく篩では、意味的にも文法的にも 99%の確率で正しい文を選択できることを明らかにした。構築されたクアジパラレルコーパスを元々の訓練データに追加して機械翻訳実験を行い、少ない訓練データの場合には、追加訓練データを３分の２まで追加した実験により、BLEU の６ポイントの非常に高い改善が得られた。通常実験では２ポイント程度の向上が得られることを明らかにした。

最後に、第２章、第３章と第４章で提案した手法を組み合わせた実験を行った。オープンアクセスでない比較的に大きい技術文書と特許のコーパスに適応し翻訳実験を行なった結果、大変厳しいベースラインより BLEU の 1.8 ポイントの改善ができた。

第５章[Conclusion and future work]では本論文をまとめ、２ヶ国語辞書の自動構築、複数語専門用語の自動抽出、パラレルコーパスの不足について考察し、今後の研究課題について論じている。

以上を要約すると、本論文では、日中・中日技術文書翻訳のための資源データ不足問題を扱い、様々な粒度で、オープンアクセスデータや訓練データを使用し、新たな資源を構築するため、自動的構築手法を提案した。様々な条件で翻訳品質向上ができたのでその有効性を明らかにしている。これらの成果は機械翻訳分野の発展に寄与するところ大である。よって、本論文は博士（工学）の学位論文として価値あるものと認める。

2017 年 6 月 22 日

審査員

主査　早稲田大学　教授　博士（工学）（グルノーブル大学）
ルパージュ・イヴ

早稲田大学　教授　博士（情報工学）（九州工業大学）古月敬之

早稲田大学　教授　博士（工学）（九州大学）　岩井原瑞穂