特許文献ニューラル機械翻訳における専門用語辞書の最適化について

Research on Dictionary Optimization Neural Machine Translation of Patent Literature

中国知識産権出版社有限責任公司

張孝飛

知識産権出版社有限責任公司・科技文献機械翻訳研究院の副院長を務め、主な研究方向は自然言語処理、機械翻訳、多言語特許データマイニングなどである。近年、国内外の重要な雑誌及び学会で３０篇以上の論文を発表し、１０個以上の国家自然科学基金、国家８６３計画、国家科学技術支持計画などの科学研究プロジェクトを完成した。機械翻訳技術の開発により、北京市科学進歩賞、国家品質検査総局「科学技術興検賞」などを受賞した。

jiqifanyi@cnipr.com　TEL +86-(0)10-8200-0860

中国知識産権出版社有限責任公司

葛昱暉

知識産権出版社有限責任公司・科技文献機械翻訳研究院の東アジア言語開発担当を務め、日・中・韓ニューラル機械翻訳の開発と特許文献データコーパスの構築に従事する。

jiqifanyi@cnipr.com　TEL +86-(0)10-8200-0860

要約

深層学習技術を取り込んだニューラル機械翻訳の登場により、機械翻訳の精度が急速に向上したものの、トレーニングと計算の複雑度は辞書語数の増加とともに激増する。逆に辞書語数をトレーニングに適する程度に削減すると、翻訳の過程において、多くの未登録語が出現し、原言語の意味が正しく表現されず、翻訳品質が低下しかねない。本稿は、日中ニューラル機械翻訳の計算効率と精度の両方の向上を目標とし、日本語特許文献の漢字語と外来語の構語ルールを分析して利用し、辞書の規模を削減しながら未登録語の割合を低く保つための専門用語辞書の最適化手法を提案することにより、特許文献の翻訳精度の向上を実現した。

キーワード：ニューラル機械翻訳、辞書最適化、特許文献機械翻訳、日中機械翻訳

1 はじめに

知識グローバル化の時代において、技術情報の交流の手段の一つである特許文献の翻訳は、経済発展と社会生活において極めて重要な役割を果たしている。一方、特許出願件数が年々増加するのに対し、コストが高いだけでなく、時間もかかる人手翻訳は、技術情報の交流を妨げるようになっている。それが原因で、人手翻訳に代わって、翻訳コストを削減できる機械翻訳エンジンの開発と普及が世界中に重要視されている。

近年以来、深層学習技術を取り込んだニューラル機械翻訳の登場により、機械翻訳の精度が急速に向上した。これまで広く使用されてきた統計的機械翻訳と比較して、ニューラル機械翻訳はモジュールフレームワークがより簡潔で学習しやすく、生データから暗黙的な特徴を直接学習し、長距離依存を捉えることができる[1]。また、訳文の完成度、論理性や言語表現習慣の面においても、統計的機械翻訳の結果より一層良い。とは言って

1　Yang Liu: Recent Advances in Neural Machine Translation, Journal of Computer Research and Development,54(6): 1144-1149, 2017

も、ニューラル機械翻訳にも欠点があり、そのうちの 1 つの顕著な欠点はデータスパースネス問題による翻訳性能の急速な低下である。ニューラル機械翻訳のトレーニングの複雑度と計算の複雑度は辞書語数の増加とともに激増するので、辞書の語数は一般に 3-8 万の間に控えるのが良いことが示されている[2]。一方、統計的機械翻訳に用いる辞書は、通常百万レベルの規模であることが知られている。つまり、ニューラル機械翻訳に適用可能にするために、その辞書から大量の語彙を除外しなければならない。これは、ほとんどの語彙がニューラル機械翻訳に適した辞書から除外されることを意味し、翻訳の過程において、多くの未登録語が出現し、原言語の意味が正しく表現されず、翻訳品質が低下しかねない。特に、Jean らの実験では、未登録語の数が語形の豊かな言語への影響がより大きいことが示されている[3]。

語形変化の豊かな日本語は、特許文献に使われる専門用語が鮮明な特徴を持っている。中国語から吸収された漢字語とカタカナで表示された多数の外来語が使用され、さらに、新語の場合、漢字語と外来語により派生する場合が多い。本稿は、日中ニューラル機械翻訳の計算効率と精度の向上を目標とし、日本語特許文献の漢字語と外来語の構語ルールを分析して利用し、辞書の規模を削減しながら未登録語の割合を低く保つための専門用語辞書の最適化手法を提案することにより、特許文献の翻訳精度の向上を実現した。

2 先行研究

2016 年にグーグルがニューラル機械翻訳を発表して以来、ニューラル機械翻訳関連の研究は急速に進んでいる。多くの学者は研究過程で辞書制限の問題がニューラル機械翻訳に与える影響を認識し、研究を行った。

Luong ら（2016）は単語 - アルファベット混合モデルを提案し、データ中の高頻度語を単語モデルを利用して処理し、低頻度語をアルファベットモデルを利用して処理することにより、未登録語が訓練に与える影響を低減する[4]。Luong らの研究により未登録語問題は一定な程度で解決できるが、形態学的な概念が重要視されていないので、語構造における意味と情報の欠落をもたらす可能性が考えられる。

これに対し、Ataman ら（2017）は言語動機に基づいた辞書を最適化する方法を開発し、ある程度で形態変化のある言語が処理中に語構造における意味と情報を欠落する問題を解決し、翻訳精度の向上を果たした[5]。しかし、Ataman らの研究は一般コーパスをめぐるものなので、新語、派生語が多い特許文献の処理に対し、不適切なところがあると考えられる。

上述の問題に鑑みて、劉ら（2018）は中国語科学技術文献のニューラル機械翻訳に存在する辞書の最適化する問題に対し、中国語科学技術用語の構語ルールに基づき、点相互情報を結合し、構造における意味と情報の欠落を回避しながら、ニューラル機械翻訳の辞書最適化をし、未登録語の数量を控えると同時に、翻訳精度の向上を実現した[6]。劉らの研究は大量の漢字語を含む日本語用語辞書の最適化に提示することはあるが、日本語の場合、外来語及び混合語も数多く存在しているので、状況により更に細かい分析が必要である。

日本語特許文献は言語の特徴から言えば、大量の漢字語及び外来語を含み、また、科学技術の迅速な発展により、新語の出現率が非常に高く、新語の構成方式も多様である。つまり、日本語特許文献のニューラル機械翻訳に使われる専門用語辞書を最適化し、翻訳精度の向上を実現するためには、漢字語と外来語との各自の処理ルール、及び語構造における意味と情報の完全性の保持の両方を考慮に入れなければならない。

2 Bin Pang: Machine Translation——From SMT to Neural Network, Digital Communication World,(12):296-297, 2016

3 Jean S, Cho K, Memisevic R, et al: On Using Very Large Target Vocabulary for Neural Machine Translation. arXiv Preprint, arXiv:1412.2007

4 Loung M, Manning C: Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models, Proc of the 54 ACL. Stroudsburg, PA: ACL:1054-1063, 2016

5 Ataman D, Negri M, Turchi M, et al: Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkey to English, Prague Bulletin of Mathematical Linguistics, 108(1):331-342, 2017

6 Qingmin Liu, Changqing Yao, Chongde Shi, Xiaojie Wen, Yueying Sun: Vocabulary Optimization of Neural Machine Translation for Scientific and Technical Document, Data Analysis and Knowledge Discovery, 3(03):76-82, 2019

3 日本語特許文献における専門用語の語彙的特徴

科学技術、及び特許文献は技術コミュニケーションの主要なアプローチの一つであり、その言語の特徴に関する研究は言語学の研究者に注目されている。国（1981）はすでに言及したが、日本語科学技術文献の中で、漢字語彙は約 47.5%を占め、カタカナで表された外来語の語彙は、混合語を含めて 16%を占めている[7]。また、近年以来、科学技術の急速な発展に伴い、新しく作られた漢字語と外来語の数量が更に激増し、特に先端科学技術を保護する特許文献において、今までより一層大きな割合を占めている。また、漢字語と外来語は多数の場合、文の主語・目的語及び補語など、名詞として使われているが、「する」を後ろにつくことで、サ変動詞として使用されることも多くなっている[8]。このような動詞は辞書に載せている可能性が少なく、また、コーパスにも滅多に出ていないので、トレーニングの精度に影響しかねない。

日本語特許文献における専門用語の語彙的特徴から見て、専門用語辞書を最適化する場合、漢字語と外来語の表記方式によって分けて、各自の構語特徴に基づいて処理するのが妥当だと考えられる。

1 漢字語の構語特徴

語源から言って、日本語の漢字語はもともと中国語に固有する語彙から引用されたものが多い[9]。長い時間に渡って、日本語特有の漢字構語素も形成し、習得された漢字語につけ、新しい語彙になって使われている。日本語特許文献の場合、既有漢字語同士の語彙複合、漢字語と日本語特有の漢字構語素との語彙複合、及び混合した語彙複合は増え続ける新しいものや新しい概念を記述するための漢字語を拡大する重要な構語手法となっている。

表 1　漢字語構語例

複合方式		例文
漢字語同士の語彙複合	漢字語＋漢字語語彙素	交通＋費 非＋常識
	漢字語＋漢字語	生物＋反応 伝統＋栽培
漢字語と日本語特有の漢字構語素との語彙複合		相関＋性 清潔＋用 一体＋化
混合した語彙複合（漢字語、日本語特有の漢字構語素の混合する場合）		大豆＋精選＋器 一体＋化＋装置 麦＋揚返＋機

表 1 が示されるように、特許文献における漢字語は、短い単語によって組み合わせて生成した長い単語（字数が三つ以上で、二つ以上の短い単語によって組み合わせて生成した単語）が多い。例えば「大豆精選器」は、「大豆」、「精選」及び「器」が組み合わせて生成した語で、このような短い単語によって組み合わせて生成した長い単語は、長さが長くなるほど、出現頻率が低くなり、このような長い単語をいくつの短い単語に分割できれば、辞書の数量を大規模に削減できる。また、その語彙は分割されているものの、ニューラル機械翻訳自体の特徴により、実際のトレーニングにおいて、各単語の間の意味的な関連性は、単語ベクトルとネットワーク構造を利用して保持することができるので、トレーニングの効果に影響を与えることはないと考えられる。

2 外来語の構語特徴

外来語は語源から言って、元来の発音を参照して英語、ドイツ語、オランダ語などの他言語から直接輸入された非日本語固有の語彙が多い。特許文献では、新語が生まれ続けているため、外来語も現代構語法に基づいて派生し続けている。劉（2013）によると、英語と同様に、日本語の特許文献において、外来語の構語は主に合成法、混成法及び接辞複合法などのいくつかの方式に依拠している[10]。

1) 合成法：二つ、またはそれ以上の既有語を組み合わせて新語を構成する方法：

7 Hongzhi Guo: The Development of Chinese Vocabulary Derivatives in Current Scientific Japanese Literature and Its Chinese Translation, Journal of Japanese Study and Research, (1):46-49, 1981

8 Chuanli Wang: Application of Loanwords in Japanese, Journal of Basic English Education, (1):33-36, 2001

9 Guowei Shen: The History of Vocabulary Exchange between Modern Japan and China, Tokyo: Kasama Shoin, 30-37: 2008

10 Chunfa Liu: Identification of the Meaning of Foreign Scientific Words in the Translation from Scientific Japanese into Chinese, Journal of University of Shanghai for Science and Technology, 35(4):300-309, 2013

表２　合成法による新語例

合成法による新語	既有語
ハイスループット high-throughput	ハイ high スルー through プット put
メタルフリー metalfree	メタル metal フリー free

2) 混成法：前に位置する既有語の語尾のいくつの音節を切り捨て、残された部分を後ろに位置する既有語と組み合わせて新語を構成する方法であり、混合によって派生された新語は一般的に二つの既有語の意味を兼ねている：

表３　混成法による新語例

混成法による新語	既有語
メディ・ケア Medi・care	メディ（カル）＋ケア Medi（cal）+care
バイオ・リズム Biorhythm	バイオ（ロジー）+リズム Bio（logy）+rhythm

3) 接辞複合法：輸入言語からの接頭辞と接尾辞を形態素として用いて新語を構成する方法：

表４　接辞複合法による新語例

接辞複合法	既有語
オート・サンプラー	Auto-sampler
デ・ガッサ	De·gasser

語構成から言うと、合成法と接辞複合法によって構成される外来語は輸入された複数の基礎外来語（または語彙素）の組合せと見なすことができ、混成法によって構成される外来語は、言語学的特徴から、前半部分を日本語特有の語彙素と見なすことも可能であるため、事実上に接辞複合法によって構成される外来語と同じように処理することができる。つまり、漢字語とほぼ同じように、長い単語をいくつの短い単語に分割し、辞書の数量を大規模に削減できる。

4 日本語特許文献の言語特徴に基づく辞書最適化方法の設計

上述した日本語特許文献の構語特徴から、特許文献の語彙は合成と派生の手法を用いて派生されたものが多く、新語を派生する各語彙素はもともとの意味を保持できることが分かった。これによって、長い単語をいくつかの短い単語に分割しても、ニューラル機械翻訳自体の特徴により、実際のトレーニングにおいて、各単語の間の意味的な関連性は、単語ベクトルとネットワーク構造を利用して保持することができる。この特性に基づいて、本稿は日本語特許文献の辞書最適化方法を設計した。その要は、高頻率の基礎語彙および特許文献における高頻率の接辞を合わせて基礎辞書を組みあて、それを用いて、多字語、低頻度語などの状況を処理することにある。

1　漢字語に対する処理

漢字語に対する最適化処理は、以下の三つのステップを含む：

1) 原始辞書の中の１～２字の語彙を保留する：

上述した原始辞書は広辞苑辞書と特許文献高頻度語を組み合わせて生成したものなので、含有された１～２字語はほぼ一番小さい意味素と考えられる。これらの１～２字語を新しい専門用語辞書の主要構成とする。

2) 特許文献語彙における３字語の高頻率の接頭辞と接尾辞を統計する：

コーパスにおける３字語の構語形式から言って、１＋１＋１、１＋２、２＋１の三つの形式があると考えられる。１＋２、２＋１の場合の３字語の首尾における語彙素を統計し、高頻度で、かつ位置情報が明確な接頭辞と接尾辞（例えば「～化」、「～用」、「非～」、「最～」など）を獲得し、１）で保留された１～２字語と合わせて、多字語、低頻度語などの状況の処理に使用する。

3) ３字語に対し、選別処理を行う：

３字語に対し、１）～２）で保留された語彙表に基づき、前向きの最大マッチング分割を行い、構語形式が１＋１＋１となる単語だけを残して、１）～２）で保留された語彙表に合わせて、新しい専門用語辞書になる。

３字以上の多字語、または低頻度語の場合、新しい専門用語辞書により、前向きの最大マッチング法を使用して分割処理を行う。これにより、辞書の規模を大幅に縮小できる。

2　外来語に対する処理

外来語に対する処理方法は漢字語と類似していて、主に低頻度語、長い単語を分割することにより辞書の最適化を図る。具体的には、以下のようになる：

1) 原始辞書における４文字以内の外来語を保留する；

４文字以内の外来語を保留するのは、特許文献において使用されている外来語の特徴を考慮したものである。日本語の特許文献で使われる外来語は、語源の原因で、西洋言語の特許文献で使われる専門用語と類似な特徴をもっている。その中で一つ顕著なのは、常用的短い単語の専用化[11]、つまり、日常生活でよく使われている短い単語に専門的意味を持たせ、その短い単語を独自で使われ、またはほかの単語と組み合わせて新語を派生することである。統計によると、このような短い単語は４文字以内のものが半数以上を占めている。また、５音節以上の単語は語構成から言うと、接辞が比較的に分割しやすいので、外来語辞書を処理する場合、４文字以内の外来語のみそのまま保留することにする。

2) ５～ 10 文字の外来語の高頻率の接頭辞と接尾辞を統計する：

多言語から直接に輸入された強い構語能力の持つ接辞以外、混成法で構成された外来語を分析して、語尾の音節が省略された前に位置する単語を見出し、接頭辞として扱い、長音節の外来語、及び低頻率の外来語の分割に用いる。

１）～２）で獲得した語・接辞集合を新しい外来語処理辞書にし、長音節の外来語、及び低頻率の外来語を分割する場合、前向きと後ろ向きの双方向最大マッチング法を使用して処理する。これにより、辞書の規模を大幅に縮小できる。

３　特殊文字に対する処理

特許文献の文構成は複雑で、漢字、ひらがなとカタカナ以外、数字、アルファベット、及び特殊文字（例えばⅨ、%、・など）が多く含まれている。これらは単語の分割に大きな影響を与える。本稿では、単語分割の易さと後処理の便利さの面から、これらの特殊文字を一般文字から分割するようにし、語意に対する影響を控える。

11 Zhini Lan, Xia Liang: A Comparative Study of Loanwords in Medical English and Japanese Vocabulary, The Science Education Article Collects, 388:160-162,2017

5　翻訳実験

辞書の最適化による翻訳精度に対する影響を検証するために、中国知識産権出版社有限責任公司が独自で開発している「科専笑飛」多戦略ニューラル機械翻訳システムを用いて翻訳実験を行った。テスト文は 2018 年～ 2019 年６月収集された特許文献コーパスから抽出された文で、未だにトレーニングされていないものである。評価は翻訳の自動評価指標である BLEU を用いた。

表５　翻訳実験データ

辞書	語彙数（万語）	BLEU	未登録語数量（日本語側）（個）	未登録語割合（日本語側）（%）
辞書 _O	45.6	29.40	346	0.77
辞書 _N	7.7	29.71	172	0.38

辞書 _O は今まで使われた専門用語辞書で、語彙量は 45.6 万語である。辞書 _N は処理後の専門用語辞書で、語彙量は 7.7 万語である。辞書 _O と比べ、辞書 _N は語彙数だけでなく、未登録語の数量もかなり減少している。翻訳精度から見て、辞書 _N の場合、BLEU 値は 1.05%向上し、翻訳精度が向上していることが分かった。また、未登録語の数量は 174 個減少し、辞書の規模を控えると同時に翻訳効果を保証する目的を実現した。

6　おわりに

本稿では、日本語特許文献に対し、言語学知識とニューラル機械翻訳特徴に基づいて、語彙の意味特徴を保つまま、専門用語辞書の規模を削減すると同時に未登録語の割合を低めに控える辞書最適化方法を設計した。また、辞書の最適化により、翻訳効果の向上も実現した。今後は、処理に難易度がより一層高いひらがなで書かれた和語、形態変化の豊かな動詞、形容詞、または特殊文字に対する処理に探索を進めていく予定である。


123456789
Dddddd
Ccccc