ニューラル機械翻訳における長文翻訳のための文分割による言い換え方法の検討

Paraphrasing Long Sentences by Sentence Segmentation for Neural Machine Translation

静岡大学情報学部講師
綱川　隆司

2008 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程単位取得退学。博士（情報理工学）。静岡大学情報学部学術研究員、同助教を経て、2019 年より静岡大学情報学部講師。自然言語処理の研究に従事。

tuna@inf.shizuoka.ac.jp　　TEL 053-478-1487

1 はじめに

特許明細書に含まれる文はしばしば一般の書き言葉に比べ長くなる傾向があり、また請求項など体言止めによる名詞句で表現する項目では最後の名詞を特徴付けるための連体修飾節や連体修飾句が非常に長いことが多くある。自然言語処理においてこのような長文の処理は長年難題として扱われており、機械翻訳でも例外ではない。モデル学習時にも長文は訓練データから取り除かれており、現状は長文を直接扱えるモデルにはなっていない。図 1 に 302 文字からなる日本語の 1 文を Google 翻訳および DeepL で英語に翻訳した例を示す。いずれの場合も、翻訳可能な長さになるようにある程度の長さの適当な部分で区切って 4 つの文に訳出していることがわかるが、単純に区切っているために訳出として適切でない部分が生じている。

極端な長文はそもそも人間にとっても可読性が低く避けられるべきであり、意味を保持しながら複数の適度な長さの文に分割したり、項目の列挙を文内に埋め込まず外部の表等のデータで示したりすることが理想的である。これらの言い換えを自動的に行い機械翻訳の入力とする研究が存在する[1]。しかし、文を単純に適切な位置で分割しただけでは文法的に正しい文にならず、意味的にも主語や接続語などを補完しないともとの意味を保持できなくなる場合がある。これらの問題を解析的に解決するには構文解析や照応解析が必要であり、長文にこれらの解析をそのまま適用すると精度が低下してしまう問題が生じる。

本稿ではこの問題に対処するため PAJ（公開特許公報英文抄録）の英訳で文分割がなされているものを日本語に逆翻訳することで言い換えデータを得るとともに、そこから機械翻訳のための長文言い換え器を構成するための方法について検討する。

【原文】
レンズ研削装置１からの研削水を貯蔵する研削水タンク１０と、研削水を濾過するための濾過タンク２２と、濾過タンク２２内に設けられて濾過タンク２２内を研削屑室Ａと濾過水室Ｂとに区画する濾過フィルタ２５とを設け、研削水タンク１０から研削屑室Ａへ研削水を送って濾過フィルタ２５透過させ、この濾過フィルタ２５を透過して濾過水室Ｂに流入した濾過水を研削水タンク１０へ研削水として戻し、研削水タンク１０から濾過水室Ｂへ研削水を送って濾過フィルタ２５を透過させ、濾過フィルタ２５を透過した濾過水を研削屑室Ａから研削水タンク１０へ研削水として戻すことによって、濾過手段の逆洗を行うレンズ研削装置の研削水処理方法及びその装置。
【Google 翻訳】（文区切りを改行して表示）
A grinding water tank 10 for storing the grinding water from the lens grinding device 1, a filtration tank 22 for filtering the grinding water, a grinding waste chamber A and a filtration water chamber provided in the filtration tank 22.
A filtering filter 25 which is divided into B and B is provided, and the grinding water is sent from the grinding water tank 10 to the grinding dust chamber A to be transmitted through the filtering filter 25.
Is returned to the grinding water tank 10 as grinding water, the grinding water is sent from the grinding water tank 10 to the filtered water chamber B to pass through the filtration filter 25, and the filtered water that has passed through the filtration filter 25 is ground from the grinding waste chamber A to the grinding water tank.
A grinding water treatment method for a lens grinding device and a device therefor, in which the filtering means is backwashed by returning the grinding water to 10.
【DeepL】（文区切りを改行して表示）
A grinding water tank 10 for storing grinding water from the lens grinder 1, a filtration tank 22 for filtering water, and a filtration filter 25 installed in the filtration tank 22 to divide the inside of the filtration tank 22 into a debris chamber A and a filtration water chamber B.
The water is fed from the grinding water tank 10 to the debris chamber A and then filtered through the filtration tank B.
The filtered water that passes through the filter 25 and flows into the filtered water chamber B is returned to the grinding water tank 10, and the filtered water from the filtered water tank 10 to the filtered water chamber B is returned to the grinding water tank 10.
The method and apparatus for treating the grinding water of a lens grinding machine for backwashing of the filtration means by returning it as water.

図 1　特許明細書に含まれる長文とその機械翻訳の例

2 関連研究

統計的機械翻訳やニューラル機械翻訳において、入力文が長くなると翻訳性能が低下することが知られている[2]。注意機構[3] や Transformer モデル[4] の採用により一般的な翻訳性能は向上しているものの、依然とし

て入力文長には制約がある。Neishi and Yoshinaga[5] は Transformer が長文を訳すときに制約となる位置エンコーディングを絶対位置でなく相対位置で行うことで性能の改善を示している。

長文を機械翻訳するための文分割については、実用的にはルールベースの手法によりある程度の長さに分割して入力する方法[6] がとられているが、元の意味を保持したまま分割するには接続詞の追加、文末表現の変更、主語の補完等の自明でない変更を加える必要が生じる。金・江原[1] は長文に対する日英機械翻訳のために文分割および主語の補完をパターンマッチングおよび統計的手法を用いて行い良好な結果を得ている。一方でニューラル機械翻訳のための文分割方法に関する研究は少数にとどまっている。

3 PAJ による文分割言い換えの獲得

NTCIR-6 PATENT データセットに含まれる 2002 年の日本公開特許公報全文から抽出した各明細書の【解決手段】の本文と、それらに対応する PAJ 中の英訳について、以下の条件に合致するものを抽出した。

- 日本語の原文が 300 文字以上で 1 文からなる。
- 英訳が 2 文以上に分割されて訳出されている。

2002 年の日本公開特許公報の 374551 件のデータから本条件に合致したのは 10372 件であった。図 2 にその一例を示す。解決手段の性質を表現するために

【原文】
レンズ研削装置１からの研削水を貯蔵する研削水タンク１０と、研削水を濾過するための濾過タンク２２と、濾過タンク２２内に設けられて濾過タンク２２内を研削屑室Ａと濾過水室Ｂとに区画する濾過フィルタ２５とを設け、研削水タンク１０から研削屑室Ａへ研削水を送って濾過フィルタ２５透過させ、この濾過フィルタ２５を透過して濾過水室Ｂに流入した濾過水を研削水タンク１０へ研削水として戻し、研削水タンク１０から濾過水室Ｂへ研削水を送って濾過フィルタ２５を透過させ、濾過フィルタ２５を透過した濾過水を研削屑室Ａから研削水タンク１０へ研削水として戻すことによって、濾過手段の逆洗を行うレンズ研削装置の研削水処理方法及びその装置。
【PAJ】
The coolant treating apparatus is provided with a coolant tank which reserves coolant for the lens grinding device, a filtering tank 22 which filters coolant and a filter 25 which separates the filtering tank 22 into a grinding dust room A and a filtered coolant room B. The coolant is fed from the coolant tank 10 to the grinding dust room A in order to be passed through the filter 25. The coolant flown into the filtered coolant room B passed through the filter 25 is returned to the coolant tank 10 as coolant. Then the coolant is fed from the coolant tank 10 to the filtered coolant room B in order to be passed through the filter 25, and filtered coolant passed through the filter 25 is returned from the coolant dust room A to the coolant tank B as coolant. The back washing for filtering means is realized by such a method and apparatus.
【PAJ の日本語訳（Google 翻訳）】
クーラント処理装置は、レンズ研削装置用のクーラントを貯留するクーラントタンクと、クーラントを濾過する濾過槽２２と、濾過槽２２を研削ダスト室Ａと濾過クーラント室Ｂとに分離するフィルタ２５とを備えている。クーラントは、クーラントタンク１０から研削ダスト室Ａに送られ、フィルタ２５を通過する。フィルタ２５を通過した濾過クーラント室Ｂに流入したクーラントは、クーラントとしてクーラントタンク１０に戻される。 そして、クーラントは、クーラントタンク１０から濾過クーラント室Ｂに送られ、フィルタ２５を通過し、フィルタ２５を通過した濾過クーラントは、クーラントダスト室Ａからクーラントタンク Ｂにクーラントとして戻される。 濾過手段の逆洗は、このような方法と装置により実現される。

図２　特許明細書原文と文分割された PAJ およびその日本語訳の例

２文以上を用いることもあるが、図２の例のように単一の文や名詞句が用いられることも少なくない。これに対し、PAJ でも単一の文として訳される場合と、この例のように複数の文に分けて訳される場合がある。この例では装置とその構成部分を最初に訳して１文とし、残りの研削水の流れと処理方法を４文で訳している。また、単純に前から順番に訳出しているとは限らず、適切な主語等を補って文法的にも正しく意味的に等価な英文を訳出している。

この PAJ の英訳を日本語に翻訳したものは、原文を英訳するために適切に文分割した言い換え表現とみなすことができる。そこで、原文と PAJ の日本語訳を並列コーパスとみなして教師ありの言い換え器を構成することで、原文を適切に文分割して英訳するための言い換え表現を得ることを考える。既存の学習モデルをそのまま用いると長文の扱いの問題が生じる。そこで、原文と言い換えに存在する共通要素を特定のトークンに置き換えることで短文化することを考える。

抽出した文分割例のうち 100 件を無作為に抽出し、文分割方法や日本語への翻訳結果について調査した。100 件の文分割数の平均は 3.75 であった。なお、本稿で調査する文分割は PAJ の作成方針によるものであり、統一的な基準や一般的なルールに基づく文分割方法ではない。

これらの 100 件のうち、PAJ において分割された各文の Google 翻訳による日本語訳が比較的良好といえるものは 54 件あった。ほとんどの場合、接続詞や接続助詞で結ばれた節をすべて分割するのではなく、適度に分割して重文または複文の形を残しているが、１文あたりの単語数が減少することで翻訳は容易になると思われる。一方、特許を構成する部品を列挙する際などに現れる長い名詞並列句は PAJ では分割されていないため長い文のままになることが多く、機械翻訳を難しくしている[1]。

２文に分割されているもので、一方が短文であり残り

1　図１の最初の文のように、「X は～である A と、～である B と、……、～である E を備える。」のような長い修飾節を含む名詞句の列挙は依存構造が複雑であり一般に訳出が難しくなる。文分割を行う例として、「X は A、B、……、および E を備える。A は～である。B は～である。……。E は～である。」のようにする方法があり、PAJ でも一部にこのように分割している例がある。

の部分が文分割されていないものが７件あり、うち３件は残りの部分の訳に明らかに不適切な部分があった。残りの部分が分割されないのは長い名詞並列句が含まれているなどの理由による。

文分割を行うために、分割した各文の主語を補ったり、原文の文末にある主題を最初の文で述べたりする例もみられた。このように原文への内容語の補完や文の並び替えを要する場合は照応解析や文脈理解に相当する処理が必要となり、データに基づく手法においては多量の学習データが必要になることが想定される。

その他、比較的細かく文分割が行われた場合においても、Google 翻訳に文分割したテキストをまとめて入力したために一部の文の出力が切り捨てられるケースがみられた。１文ずつ入力すればこの問題は改善されるが、訳語の一貫性は低下するおそれがある。

4 長文を文分割するための言い換え器の構築方法の検討

以上の分析をもとに、原文と PAJ の日本語訳を言い換え器構築のための並列コーパスとして用いるために、Luong et al.[7] による方法に従い下記の処理を行うことを考える。

- 原文と PAJ の日本語訳を形態素解析して単語に区切り、それぞれに含まれる各単語の対応関係を得る。
- 語順や文の順序に大きな変化があるデータ、および対応のない語が多く現れるデータを並列コーパスから除外する。
- 原文と PAJ の日本語訳の各単語のうち、未知語であるものと、サブワード化によって分割される単語をともに未知語を表す特殊なトークン（[UNK]）に置換し、さらに[UNK]が連続する部分を単一の[UNK]に置換する。
- 言い換え器による言い換えの出力時には、入力時に[UNK]に置き換えられた部分をもとの語句に復元する。

図３に図２の原文と PAJ の日本語訳に含まれる未知語を[UNK]に変換する例を示す。ここで、原文と PAJ の日本語訳のそれぞれに含まれる[UNK]の間には Luong et al. [7] による方法に従い対応関係を別途付けているものとする。

本方法による言い換え器の構築には、入力するテキストを訓練可能な程度に短くする必要があり、そのために未知語として扱わない語彙数や[UNK]に変換すべき語彙を調整しなければならない。また、除外対象となるデータの抽出方法、原文と PAJ 日本語訳の[UNK]間の対応関係の獲得、言い換え時の[UNK]の復元方法の検討が今後の検討課題である。

【原文（未知語の列を[UNK]に置換した例）】
[UNK]からの[UNK]を貯蔵する[UNK]と、[UNK]と、[UNK]内に設けられて[UNK]内を[UNK]と[UNK]とに区画する[UNK]とを設け、[UNK]から[UNK]へ[UNK]を送って[UNK]、この[UNK]に流入した[UNK]を[UNK]へ[UNK]として戻し、[UNK]から[UNK]へ[UNK]を送って[UNK]させ、[UNK]を[UNK]から[UNK]へ[UNK]として戻すことによって、[UNK]処理方法及びその装置。
【PAJ の日本語訳（未知語の列を[UNK]に置換した例）】
[UNK]処理装置は、[UNK]装置用の[UNK]と、[UNK]と、[UNK]とを備えている。[UNK]は、[UNK]から[UNK]に送られ、フィルタ２５を通過する。フィルタ２５を通過した[UNK]に流入した[UNK]は、[UNK]として[UNK]に戻される。 そして、[UNK]は、[UNK]から[UNK]に送られ、フィルタ２５を通過し、フィルタ２５を通過した[UNK]は、[UNK]から[UNK]に[UNK]として戻される。 [UNK]は、このような方法と装置により実現される。

図３　未知語の列を[UNK]に置換した例

参考文献

[1] 金淵培、江原暉将：日英機械翻訳のための日本語長文自動短文分割と主語の補完、情報処理学会論文誌、Vol. 35, No.6, pp. 1018-1028. (1994).

[2] Koehn P. and Knowles, R.: Six Challenges for Neural Machine Translation, In Proc. of the 1st Workshop on Neural Machine Translation, pp. 28-39. (2017).

[3] Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, ICLR 2015, (2015).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, NIPS 2017, pp. 5998-6008. (2017).

[5] Neishi, M. and Yoshinaga, N.: On the Relation between Position Information and Sentence Length in Neural Machine Translation, In Proc. of the 23rd Conference on Computational Natural Language Learning, pp. 328-338. (2019).

[6] 園尾聡：ニューラル機械翻訳による特許機械翻訳システムの開発、Japio YEAR BOOK 2019, pp. 296-301. (2019).

[7] Luong, M.-T., Sutskever, H., Le, Q.V., Vinyals, O., and Zaremba, W.: Addressing the Rare Word Problem in Neural Machine Translation,

ACL-IJCNLP 2015, pp. 11-19. (2015).