全文を翻訳しようとしない機械翻訳
—ワードグラフによる部分的機械翻訳の試み—

Partial Machine Translation by Word Graph Generated from Parallel Corpus

長岡技術科学大学電気系准教授 **山本　和英**

PROFILE: 豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士（工学）。1996年～2005年（株）国際電気通信基礎技術研究所（ATR）、2002年～現在まで長岡技術科学大学。2012年から電子情報通信学会言語理解とコミュニケーション（NLC）研究会委員長。自然言語処理、及びテキストマイニングの研究に従事。

1 「要約翻訳」タスクの提案

機械翻訳の研究が始まって以来、機械翻訳の仕事は入力された原言語表現を如何にして正確に目的言語表現に変換するかであって、これに疑問の余地は全くなかった。どのような翻訳手法を採用するかに関係なく、入力表現の全単語が何らかの形で出力表現に反映され（単語によっては翻訳の必要がないとして無視され）、翻訳が行われる。近年の統計的機械翻訳においても、この前提は何も変わらない。

ところが、我々人間は必ずしもこのような翻訳を行っていない。特に個人的に通訳する（あるいは何か伝えたいことを外国語で話す）際はこれが顕著である。翻訳しにくい表現や意味の分からない表現は無視し、特に相手に伝えなくてもよさそうな内容は省略する。職業として翻訳・通訳する場合などはこの限りではないが、いずれにしてもこのような「適当に要約しながら翻訳」する場面は確かに存在する。

我々は以上の着眼点を持って機械翻訳モデルの検討を行った。これにはいくつかの理由がある。まず、（1）現在では機械翻訳モデルそのものの研究があまり行われていないことが問題だと考えた。機械翻訳の研究は今も少なくないが、その多くは統計的機械翻訳に関する改良もしくはチューニングの類と言えるもの、あるいは関連する言語資源の（自動・手動）構築で、新しい機械翻訳モデルを提案しようという動きはほとんど見られない。次に、（2）入力文の全単語を必ず翻訳しなければならない、という制約を緩めたタスクを設定することで、従来よりも柔軟な翻訳手法が可能になるのではないかと考えた。また、前述した通り（3）人間がしばしば行う作業に近く、さらに（4）このような翻訳作業は実際に需要がある。

本稿では、原言語文と原言語文の一部の表現（部分的表現）を入力として、部分的表現を中心とした翻訳を行う部分的機械翻訳の試み[1]を紹介する。翻訳する部分を限定することにより出力結果の情報を簡潔化し、翻訳精度と可読性を持った機械翻訳を目指す。

2 部分的機械翻訳

本稿で研究の対象とする部分的機械翻訳とは、原言語文と原言語文の一部の表現（部分的表現）が入力されたとき、部分的表現を中心とした翻訳を出力する。例えば「This product provides your skin with moisture and keeps it healthy.」という原言語文があったとき、部分的表現「This product keeps」と共に部分的機械翻訳器に入力すると「肌をすこやかに保つ」と出力される。

部分的表現は自動的に抽出することができれば、どのような形式でも受付ける。例えば入力文の中からユーザーが指定した単語列や、入力文を構文解析して語を組み合わせたSVO型のようなパターンでも良い。

部分的表現自体の翻訳ではなく、部分的表現を中心とした翻訳とする理由は、本研究が述べる部分的機械翻訳器では入力文がはじめにあり、入力文の文脈の存在を前提とした翻訳結果になることを期待している

ためである。例えば、「ProductA is a hair dye for women that dyes with foam, giving you an easy and even color in your own home.」を部分的表現「ProductA is a hair dye」で翻訳した場合、入力文の対訳は「ProductA は泡で染めるから、色ムラもなく自宅染めが簡単にできる女性用ヘアカラー（おしゃれ染め）です」なので、「hair dye」は「ヘアカラー」もしくは「おしゃれ染め」と訳される。一方、同じ部分的表現「ProductA is a hair dye」と「ProductA is a cream-type men's hair dye for gray hair, that spreads well without dripping.」という入力があった場合、対訳は「ProductA は、のびがよくたれにくいクリームタイプの男性用白髪染めです」なので、「hair dye」は「白髪染め」と訳される。

部分的表現のみ入力した場合、このような訳し分けを行うことができない。また、入力文を文脈として部分的表現を翻訳する場合、自然な翻訳のために部分的表現のみでは読み取れない情報が付与されることがある。そのため、本稿では部分的表現を中心とした翻訳としている。

3 関連研究

本研究に関連した手法として、Filippova の研究 [2] を紹介する。Filippova の手法は機械要約の中でも文圧縮と呼ばれるタスクに属する。Filippova の文圧縮手法は同じ内容を取り扱った新聞記事を収集し、記事の内容を一文で説明する。記事集合の一文目を利用して、文内の語をノード、語同士の繋がりをエッジとしたワードグラフと呼ばれる有向グラフを生成する。エッジには文書集合を説明する度合いを重みとして付与している。この手法は記事の一文目という不要な情報を含む文集合からでも記事を説明する文が生成できる。

我々の提案手法では同じ部分的表現を持つ対訳文集合を抽出し、対訳文集合内で Filippova の文圧縮手法を応用する。部分的表現で抽出された対訳文集合は Filippova の文圧縮手法と同様に、入力文にとって不要な情報を含む。本手法では、与えられる部分的表現が自動的に抽出でき、対訳文集合を抽出できれば言語に依存しない手法となっている。

4 手法の概要

本稿が提案する部分的機械翻訳の流れを以下に示す。本稿では英日翻訳を対象としているが、与えられる部分的表現が自動的に抽出でき、対訳文集合を抽出できれば言語に依存しない手法である。

1. 翻訳したい英文と部分的表現が入力される
2. 対訳コーパスから英語側の文に同じ部分的表現を持つ対訳対を抽出し、日本語側の文を対訳文集合とする
3. 対訳文集合各文の単語をノード、語と語の繋がりをエッジとしたワードグラフを作成
4. ノードの重要度や文脈に関する重みを付与
5. ワードグラフから短い順に k 個の経路を得る
6. k 個の経路をリランキングし、1 位となった経路を部分的機械翻訳として出力

4.1 ワードグラフの作成

ワードグラフは文集合内の語をノードとし、語の繋がりをエッジとした有向グラフである。文集合には文頭を表す「START（S）」ノードと、文末を表す「END（E）」ノードが含まれている。

本手法ではこのワードグラフの START から END までの経路が部分的翻訳として適切になるよう、経路として適切なエッジは重みが小さくなるように重み付けを行い、経路内のエッジの重みの合計を距離として最短経路を探索することで翻訳を出力する。

例として、以下の３文からワードグラフを作成した結果を図１に示す。

（例文）
文１：肌 を 柔軟 に 保つ 化粧 水 です
文２：肌 を きれい で すこやか に 保ち ます
文３：肌 を 健康 的 に 保つ 化粧 水 です

ワードグラフから最短距離となるパスを選択する本手法では少ない数のノードを持つ経路を選択する傾向にある。１ノードを１語として扱うと、定型表現などの１語で表せない表現が複数ノードになり、複数ノードによる

表現を選択しなくなるという問題がある。

このワードグラフを用いて「This product provides your skin with moisture and keeps it healthy.」を翻訳する場合、入力文に「柔軟」に対応する語が無いので、「柔軟 - に」の経路ではなく「を - 健康 - 的」の経路を通過することが望ましい。本手法では「を - 柔軟」間の重みよりも「を - 健康」間の重みの方が小さくなるようにエッジの重み付けを行うが、経路上のエッジの数が増える程、経路全体としての距離が長くなるため「を - 健康 - 的 - に」という経路よりも「を - 柔軟 - に」を通過する経路の方がエッジの数が少なくなり、「を - 柔軟 - に」の経路の方が有利になる可能性がある。そのため、このような語同士を結合して１ノードとする。

本稿ではノードの結合のために「エッジの出入りが１本のノード同士を結合」という手法と「文集合により動的にノードを結合」という２種類の手法を比較する。前者の手法ではまず 25 語のストップワードを設定し、これらの語を含むノードを結合する（図２）。この結果エッジの出入りが共に１本のノードに対して結合処理を行う（図３）。後者の方法は、ストップワードや重要と考えられる語ををワードグラフ毎に動的に決定し、それらの語に繋がる語とのみ接続する。語の決定は、最頻出語に対してある一定割合以上の頻度を持つ語とした。

４.２ エッジの重み

エッジに付与する重みは、ワードグラフ内でのエッジの重要度 ScoreF と入力文の文脈をワードグラフの各エッジの重みへ反映させるスコア ScoreT の重みつき線形和によって計算する。重要エッジスコア ScoreF はワードグラフ内でのエッジの重要度を表し、重要度の高いエッジを通ることを期待されている。重要なエッジとはワードグラフ内での出現頻度の高いエッジを表す。同じ表現によって収集された文集合内には共通の表現が含まれており、ワードグラフ内での出現頻度が高いと予

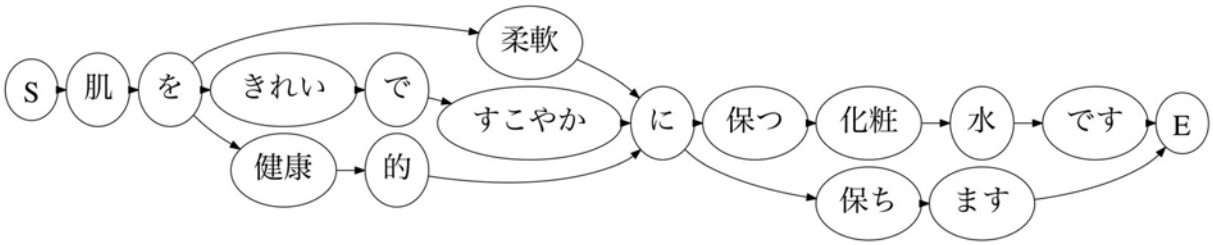

図１改良前のワードグラフ

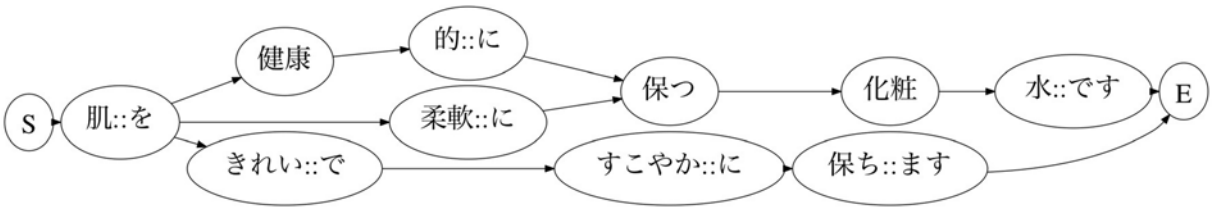

図２　ストップワード結合後のワードグラフ

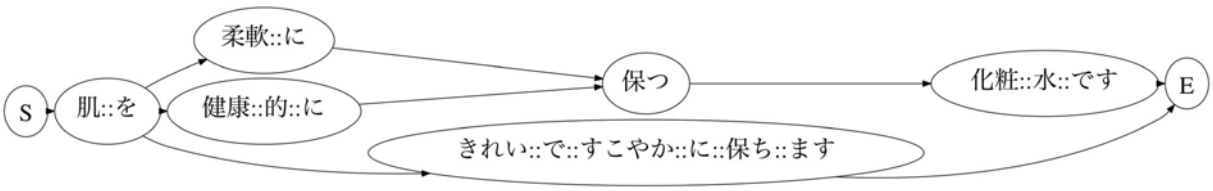

図３　エッジの出入りが１本のノードを結合後のワードグラフ

想できる。単語対応確率スコア ScoreT は入力文の文脈をワードグラフの各エッジの重みへ反映させるスコアである。このスコアは統計的機械翻訳で算出する単語対応確率を用いた。

5 手法の評価

評価実験では英日翻訳による評価実験において翻訳結果に間違った内容を含んでいる割合を６段階、流暢さを５段階で評価し、指定した部分の翻訳精度についても評価した。

既存の文圧縮手法でのエッジの重み計算を重要ノードスコアと単語対応確率スコアに置き換えた手法では翻訳精度が低く、ノードの結合方法の変更や部分的表現に対するノードの通過の強制等の処理を追加することで、翻訳結果に平均 72.6%の正しい情報を含む結果となった。また、流暢性の高い翻訳が出力され、指定した部分の翻訳精度は 78%となった。また、入力文全体を正しく翻訳した場合に対する文圧縮率は、文集合により動的にノードを結合した手法で平均 48.9%となっており、半分程度にまで短い出力となっていた。

6 おわりに

本稿の冒頭で述べた通り、機械翻訳の研究は統計的機械翻訳が主流となっているが、より研究の幅を広めるためには統計的機械翻訳以外の様々な機械翻訳モデルを模索していく必要性を感じている。本稿で紹介したものはその試行例であるが、今後も様々な観点から機械翻訳モデルの検討を進めていきたい。

使用したツール

(1) IBM 翻訳モデル構築ツール「GIZA++」, 2001-01-30, http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

(2) 形態素解析器MeCab, Ver.0.98, http://mecab.sourceforge.net/

(3) 英文構文解析器 Stanford Parser, Ver. 2.0.2, http://nlp.stanford.edu/software/lex-parser.shtml

参考文献

[1] 井手上 雅迪 . 対訳コーパスから生成したワードグラフによる部分的機械翻訳 . 長岡技術科学大学修士論文 2013, https://dl.dropboxusercontent.com/u/2152477/arc/13/13thesis-ideue.pdf

[2] Katja Filippova, Multi-sentence compression: Finding shortest paths in word graphs. Proceedings of International Conference on Computational Linguistics（Coling 2010）, pp. 322-330, 2010