英語を中間言語とする露日NMTの構築

Russian to Japanese NMT system using English as a pivot language

元山梨英和大学教授
江原　暉将

1967年早稲田大学理工学部卒。同年NHK入局。2003年諏訪東京理科大学教授。2009年山梨英和大学教授。2015年退職。アジア太平洋機械翻訳協会（AAMT）／Japio特許翻訳研究会委員。

有限会社アジア産業　研究開発部部長
岡　俊行

1983年東京工業大学数学科卒。株式会社クロスランゲージなどを経て、現在アジア産業に拠点を置きつつ、主にプログラマとして活動中。

1 はじめに

ニューラル方式機械翻訳（NMT：Neural Machine Translation）によって機械翻訳の性能が格段に向上し、特許翻訳においても実用が進んでいる。しかし、このような機械学習による翻訳システムの場合、大量の訓練データを必要とし、言語対によっては訓練データが不足する場合がある。このような場合は、低リソース言語対とよばれ、克服する方法がいくつか提案されている（例えば文献［1］参照）。

その一つとして中間言語を用いる方式がある。例えば言語Aと言語Bの間の対訳データが少ない場合でも、間に言語Cを挟むことでAとCの対訳データおよびCとBの対訳データは共に大量に存在する場合がある。例えば言語Cとして英語を挟むことが考えられる。このような方式は中間言語方式と呼ばれ言語Cは中間言語あるいはピボット言語と呼ばれる。

特許文献の場合、低リソース言語対の例としてロシア語から日本語への翻訳がある。露日の対訳データは少量でも、露英および英日の対訳データは大量にある。そこで中間言語方式で翻訳システムを構築することが考えられる。

2 カスケード方式

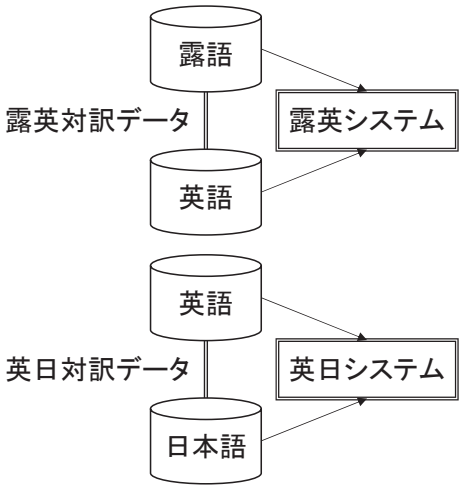
中間言語方式の中で最も単純な方式はカスケード方式である。中間言語として英語を採用した露日翻訳システムの場合、カスケード方式は【図1】のようになる。

訓練時には、露英対訳データから露英翻訳システムが構築され、英日対訳データから英日翻訳システムが構築される。露英対訳データの英語部分と英日対訳データの英語部分は独立であり関係がなくても良い。

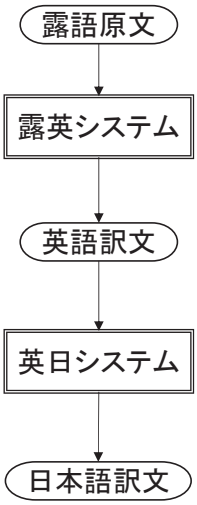
翻訳時には露語原文を露英システムで英語に翻訳し、翻訳結果の英語訳文を英日システムに入力して日本語訳文を得る。このように英語を中間言語として挟み、露英システムと英日システムを直列（カスケード）に接続して露日翻訳システムとするのがカスケード方式である。

3 順方向翻訳方式

カスケード方式の露英対訳データの英語部分を何らかの翻訳システムで日本語に翻訳する。この英語部分と日本語への翻訳結果（疑似日本語と呼ぶ）とを元々の英日対訳データに加えることで英日の対訳データ量を増加できる。この拡張した英日対訳データから英日システムを



(a) 訓練時



(b) 翻訳時

図１　カスケード方式

訓練することで英日システムの精度が向上できる可能性がある。このような方式を順方向翻訳方式（forward translation method）と呼ぶ。訓練時の処理手順を【図 2】に示す。翻訳時の処理手順は【図 1】の（b）と同じである。

4 逆方向翻訳方式

順方向翻訳方式とは逆にカスケード方式の英日対訳データの英語部分を何らかの翻訳システムで露語に翻訳する。この英語部分と露語への翻訳結果（疑似露語と呼ぶ）とを元々の露英対訳データに加えることで露英の対訳データ量を増加できる。この拡張した露英対訳データから露英システムを訓練することで露英システムの精度が向上できる可能性がある。このような方式を逆方向翻訳方式（back translation method）と呼ぶ。訓練時の処理手順を【図 3】に示す。翻訳時の処理手順は【図 1】の（b）と同じである。

5 順および逆方向翻訳方式

順方向翻訳方式と逆方向翻訳方式を組み合わせるのが順および逆方向翻訳方式（forward and back translation method）である。本方式には、２種類が考えられる。

第一の方式は、順方向翻訳方式と逆方向翻訳方式を単純に組み合わせる方式であり、逆方向翻訳方式で得られ

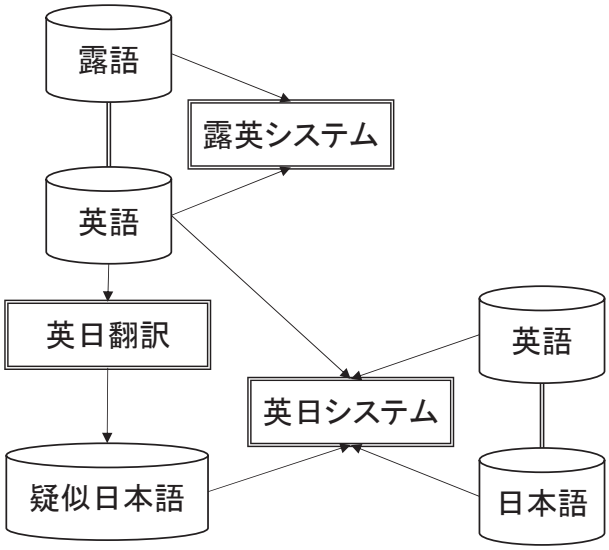


図２　順方向翻訳方式

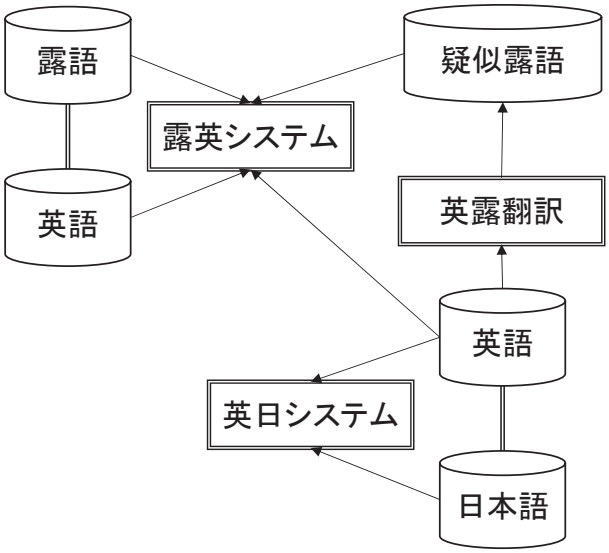


図３　逆方向翻訳方式

た露英システムと順方向翻訳方式で得られた英日システムとを直列（カスケード）に接続する方式である。本方式を順および逆方向翻訳方式（その 1）と呼ぶ。本方式の訓練時の処理手順を【図 4】に示す。翻訳時の処理手順は【図 1】の（b）と同じである。

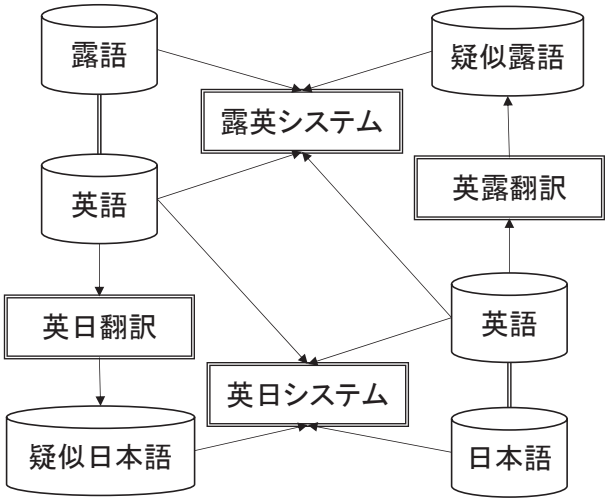


図 4　順および逆方向翻訳方式（その 1）

第二の方式は、順方向翻訳方式で得られた英日対訳データの日本語部分と、逆方向翻訳方式で得られた露英対訳データの露語部分を直接結び付けて露日システムとする方式であり、中間言語の英語は陽には現れない。本方式を順および逆方向翻訳方式（その 2）と呼ぶ。本方式の訓練時の処理手順を【図 5】に示す。翻訳時の処理手順は【図 1】の（b）とは異なり、露語を直接日本語に翻訳するものとなる。

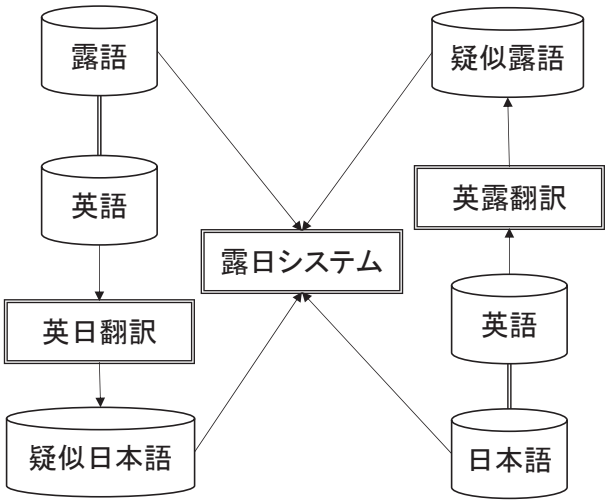


図 5　順および逆方向翻訳方式（その 2）

6 実験設定および結果

実験に用いたデータは、パテントファミリーから得られた公開特許公報（publication of unexamined patent applications）の詳細説明（description）部分を用いた。露英、英日、露日の各対訳データに対する訓練、開発、試験データの規模（文対数）を【表 1】に示す。

表 1　実験に用いた対訳データの規模

対訳データ	訓練	開発	試験
露英	161 万	1,616	---
英日	800 万	2,928	---
露日	---	---	17,874

表 1 に示した訓練データを元に、順方向翻訳（英日翻訳）や逆方向翻訳（英露翻訳）を用いてデータ規模を拡張する。その際、翻訳ツールとして何を用いるかが問題となる。今回の実験では以下のようなツールを用いた。順方向翻訳（英日翻訳）ツールは英日の訓練データ（800 万文対）で訓練された NMT システムを用いた。一方、逆方向翻訳（英露翻訳）ツールは市販の NMT ツール（Promt NMT 21 [2]）を用いた。英露の訓練データの規模（161 万文対）が比較的少量であるためである。

また、単語分割には、露語と英語は Moses の tokenizer.perl [3] を用いた。日本語は unidic [4] ベースの Mecab [5] を用いた。露語と英語は独自ツールによって小文字化処理を行った。このようにして訓練データを拡張した結果、各種方式の訓練データの規模（文対数）は【表 2】のようになった。

表 2　各方式の訓練データの規模

方式	露英	英日	露日
カスケード方式	161 万	800 万	---
順方向翻訳方式	161 万	961 万	---
逆方向翻訳方式	961 万	800 万	---
順および逆方向翻訳方式（その 1）	961 万	961 万	---
順および逆方向翻訳方式（その 2）	---	---	961 万

実験に用いた翻訳処理系は Marian Transformer [6] である。Marian での処理の前に独自ツールによるサブワード化および SentencePiece [7] による分割を行っ

た。SentencePiece のモデルは露語・英語・日本語あわせて 3 万ピースを用い、訓練データから訓練した。モデルタイプは unigram である。

各種方式に対して露日翻訳結果の BLEU 値 [8] を測定した。結果を【表 3】に示す。

表 3　各種方式の露日翻訳結果の BLEU 値

方式	BLEU
カスケード方式	32.57
順方向翻訳方式	33.08
逆方向翻訳方式	32.86
順および逆方向翻訳方式（その 1）	33.41
順および逆方向翻訳方式（その 2）	33.63

【表 3】から以下のことが言える。

・データ拡張を行う方式はカスケード方式より BLEU 値が高い。

・順方向翻訳方式は逆方向翻訳方式より BLEU 値が高い。

・順および逆方向翻訳方式は順方向あるいは逆方向単独の方式より BLEU 値が高い。

・順および逆方向翻訳方式の中では、（その 2）の方が（その 1）より BLEU 値が高い。

最後の事項は、（その 2）の方法が中間言語である英語を陽に用いていないことに起因するのではないかと思う。

7 まとめ

英語を中間言語とする露日 NMT について各種データ拡張方式を実験し評価した。その結果、順および逆方向翻訳方式（その 2）が最も BLEU 値が高かった。

本文と類似の実験が文献［9］でも成されている。文献［9］での対象分野は報道分野であり、特許分野とは異なるものの結果は類似したものとなっている。

本文で述べた方式は対訳コーパスだけでなく単言語コーパスを用いても実行できる。大規模な単言語コーパスを用いた実験は今後の課題である。

参考文献

[1] Rui Wang , Xu Tan , Renqian Luo , Tao Qin, and Tie-Yan Liu : A Survey on Low-Resource Neural Machine Translation, *arXiv:2107.04239*, 9 July 2021.

[2] PROMT Company : PROMT Master NMT 21 English ⇔ Russian, *https://www.promt.com/translation_software/home/forwindows/promt-master-neural-ere/*, 2021 年 8 月アクセス。

[3] Philipp Koehn : Moses, Statistical Machine Translation System, User Manual and Code Guide, *http://www.statmt.org/moses/manual/manual.pdf*, 2021 年 8 月アクセス。

[4] 国立国語研究所 : Unidic, *https://unidic.ninjal.ac.jp/*, 2021 年 8 月アクセス。

[5] 工藤 拓 : MeCab: Yet Another Part-of-Speech and Morphological Analyzer, *https://taku910.github.io/mecab/*, 2021 年 8 月アクセス。

[6] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andre F.T. Martins, and Alexandra Birch : Marian: Fast Neural Machine Translation in C++, *Proceedings of ACL 2018, System Demonstrations*, pp.116–121, 2018.

[7] Google : SentencePiece, *https://github.com/google/sentencepiece*, 2021 年 8 月アクセス。

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu : BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, July 2002.

[9] 美野 秀弥 , 衣川 和尭 , 伊藤 均 , 後藤 功雄 , 山田 一郎 , 田中 英輝 , 川上 貴之 , 大嶋 聖一 , 朝賀 英裕 : 単言語データを用いた逆翻訳と順翻訳によるデータ拡張の効果の比較 , 言語処理学会 第 *27* 回年次大会 発表論文集 , pp.118-122, March 2021.