逆翻訳による高品質な大規模擬似対訳コーパスの作成

Imankulova Aizhan　佐藤 貴之　　小町　守
首都大学東京
a.t.imankulova@gmail.com, sato-takayuki@ed.tmu.ac.jp, komachi@tmu.ac.jp

1　はじめに

大規模対訳コーパスは，統計的機械翻訳 (PBSMT) やニューラル機械翻訳 (NMT) のモデル学習において不可欠な言語資源である．これらの機械翻訳の精度は，対訳コーパスの量と質に大きく依存する．質の高い大規模対訳コーパスを作成するには，大量のテキストに対して，専門家の人手による翻訳を要する．その結果，現存する大規模対訳コーパスの多くは，言語とドメインが限られている．一方で，ほぼ全ての言語において，大規模な単言語コーパスは利用可能である．

そのため，単言語コーパスから擬似対訳コーパスを作成する研究が行われている．PBSMT では Bond ら [1] は，語順や軽微な語彙のバリエーションを考慮して原言語側の文を言い換える手法を提案した．言い換えはコーパスの原言語側に追加され，対応する目的言語側の文が複製される．NMT では Zhang ら [11] は原言語側の単言語コーパスとその機械翻訳文による擬似対訳コーパスを生成する手法を提案した．Sennrich ら [10] は，目的言語側の単言語コーパスの文を原言語の文に機械翻訳し擬似コーパスを得て，元の対訳コーパスと擬似対訳コーパスを合わせた学習コーパスで NMT モデルを再学習することにより，精度を大きく向上させた．しかし，逆翻訳した文を全て学習に用いるため，学習を妨げるような質の低い翻訳文が含まれうるという問題点がある．

そこで，本研究では，単言語コーパスと機械翻訳によって作られた擬似コーパスだけを使うことにより，対訳コーパスを持っていなくても，翻訳モデルを学習可能であることを示す．さらに，先行研究では学習データをランダムで選択していたが， 本研究では，sentence-level BLEU+1 (以下 BLEU+1) [8] を用いて，作成した擬似対訳コーパスの文精選を行う．これにより，擬似対訳コーパスに含まれるノイズが取り除かれ，より良い擬似対訳コーパスを得ることが可能であることを示す．ロシア語-日本語の小規模な言語対の機械翻訳に対して有効な手法の一つとして考えられて

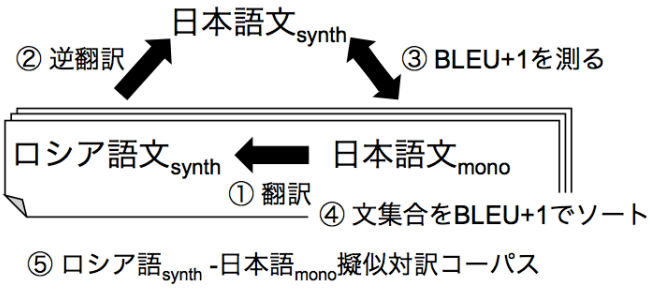

図 1: 露日翻訳における擬似対訳コーパスの作成の流れ

いるピボット翻訳手法と提案手法である擬似対訳コーパスによる翻訳結果を比較した結果，BLEU が+13 ポイント向上した．単言語コーパスから作られた擬似対訳コーパスを精選することで BLEU が+3 ポイント向上した．この結果から，擬似対訳コーパスに対して精選する手法が有効であることが示された．

2　先行研究

これまでに，小規模な言語対における機械翻訳に対して，PBSMT を用いた複数の手法が考案されてきた．特定の言語対で十分な対訳コーパスが得られない場合，中間言語を用いたピボット翻訳が有効な手段として知られている．ピボット翻訳では，原言語 A から目的言語 B への翻訳の際，言語 A からピボット言語 P に変換し，その後，言語 P から言語 B へ翻訳する．ピボット翻訳では中間言語を含む二つの翻訳モデルを合成するなど，ピボット翻訳に特有な操作によって高い翻訳精度が得られることが知られている．[3]

Cheng ら [2] は，ピボット言語として英語を用い，ドイツ語-フランス語，およびスペイン語-フランス語の言語対で，ニューラルピボット翻訳に取り組んだ．原言語-目的言語の 100k 文対の対訳コーパスを学習に加えることで，原言語から目的言語への方向だけでなく，原言語からピボット言語へ，およびピボット言語から目的言語への方向で大幅な改善を達成した．

表 1: Tatoeba Project のデータセット

コーパス	Ru-En	En-Ja	Ru-Ja
Train	95,000	95,000	10,000
Dev	500	500	500
Test	500	500	500

また，Zoph ら [13] は転移学習を用いた NMT の手法を提案した．大規模な対訳コーパスが存在する言語対で事前学習を行った後，学習したモデルを各パラメータの初期値として，小規模な対訳コーパスの言語対で学習する．この手法による，小規模な言語対における翻訳精度の向上が報告されている．

また，複数の言語対を用いる NMT の手法が提案されている．Dong ら [4]，Zoph ら [12]，Firat ら [5] は，原言語，目的言語の種類に応じて，それぞれ Encoder，Decoder を割り当て，資源が大規模な言語対の学習が小規模な言語対の精度向上に貢献することを示した．同じく，Firat ら [6] はあらかじめ訓練された多方向多言語モデルを用いて後でモデルによって生成された疑似対訳コーパスで微調整することで“ゼロリソース”翻訳を行った．Johnson ら [7] の GNMT Zero-Shot という手法では 8 層の Encoder と 8 層の Decoder により，複数の言語対で学習して，未学習の言語対を翻訳することを可能にした．

本研究では，上記の手法と異なり，他言語と大規模な計算リソースを用いない，直接翻訳を行うためのシンプルな手法を提案した．

3 擬似対訳コーパスの作成

本研究では，単言語コーパスを用いた擬似対訳コーパスを作成する手法について示す．

図 1 のように提案手法の手順は以下の通りである:

1. 単言語コーパスを他言語に機械翻 訳し，擬似原言語側コーパスを獲得する．ここで先行研究 [10] のように精選なしの擬似対訳コーパスが得られる．
2. 擬似原言語側コーパスを機械翻訳し，擬似目的言語側コーパスを獲得する．
3. もとの単言語コーパスを参照訳として，擬似目的言語側コーパスの BLEU+1 を測る．
4. 擬似原言語側コーパスと対応する目的言語側の単言語コーパスの文を BLEU+1 が高かった順にソートする．
5. スコアの高かった文対から順に擬似原言語側コーパスの文を原言語側のコーパスとし，目的言語側の単言語コーパスの文を目的言語側のコーパスとして扱う．得られたコーパスを提案手法の精選ありの擬似対訳コーパスとする．

4 擬似対訳コーパスを用いたロシア語-日本語翻訳実験

4.1 実験設定

原言語としてロシア語，目的言語として日本語を用いる．比較手法のピボット翻訳におけるピボット言語は英語とする．

本研究では，擬似対訳コーパスの作成に必要なロシア語・日本語の翻訳のために Translate Shell[1] から PBSMT である Google Translate を用いる．

訓練コーパスからの機械翻訳の学習には，PBSMT システムとして Moses[2] を，NMT システムは自ら実装したシステム[3] を用いた．BLEU+1 は mteval Toolkit[4] の mteval-sentence を用いて測定した．ロシア語と英語の文に対し，Moses の添付スクリプトを用いて，トークナイズ，正規化を行った．日本語文の分かち書きには MeCab 0.996 と IPAdic 辞書を用いた[5]．また，訓練時には 40 単語以上の文を排除した．翻訳結果の比較には BLEU [9] を用いた．

4.2 データセット

本実験で用いる対訳コーパスは，Tatoeba Project[6] から抽出した．表 1 のように，ピボット翻訳の実験において用いるデータ日本語-英語は 95k 文対，ロシア語-英語は 95k 文対である．ロシア語から日本語の直接翻訳に用いるデータは 10k 文対である．同じドメインで提案手法の実験を行うために日本語の単言語コーパスとして Tatoeba Project から 95k 文を抽出した．

大規模な日本語の単言語コーパスとしては BCCWJ[7] を用いる．BCCWJ の日本語の単言語コーパスから前処理の結果で 2,355,503 文を取得した．

対訳コーパスを用いた機械翻訳では文数が増加するにつれ翻訳精度が上がる．しかし，擬似対訳コーパスでは対訳コーパスと違いノイズが含まれている．そのため，擬似対訳コーパスを用いた機械翻訳では，文数を増やしても翻訳精度が必ずしも上がるとは限らない．

[1] https://github.com/soimort/translate-shell
[2] https://github.com/moses-smt/mosesdecoder
[3] https://github.com/tmu-nlp/NMT2016
[4] https://github.com/odashi/mteval
[5] http://taku910.github.io/mecab/
[6] https://tatoeba.org/jpn/
[7] http://pj.ninjal.ac.jp/corpus_center/bccwj/

表 2: Ru-Ja 言語対で BCCWJ からの擬似対訳コーパスのみと対訳コーパスも用いた機械翻訳の BLEU

文対	PBSMT				NMT			
	擬似対訳コーパスのみ		擬似対訳+対訳コーパス		擬似対訳コーパスのみ		擬似対訳+対訳コーパス	
	精選なし	精選あり	精選なし	精選あり	精選なし	精選あり	精選なし	精選あり
10k	5.53	**5.83**	-	-	1.59	**2.09**	-	-
50k	9.65	**11.80**	21.22	**21.96**	3.18	**5.14**	8.42	**10.65**
100k	11.48	**14.55**	22.33	**23.42**	3.74	**7.92**	8.89	**12.12**
500k	15.98	**17.14**	23.89	**23.99**	8.22	**11.47**	11.08	**12.97**
1M	**16.25**	15.67	23.86	**25.21**	9.54	**11.07**	12.02	**13.15**
2M	**15.93**	15.81	22.42	**24.38**	10.58	**11.09**	**10.87**	10.74

表 3: ロシア語-日本語言語対でピボット手法と比較

手法	文対	PBSMT	NMT
Tatoeba ピボット (ベースライン)	95k	11.51	11.10
BCCWJ 精選なし	95k	12.35	3.43
BCCWJ 精選あり	95k	14.38	6.78
BCCWJ 精選なし + Tatoeba 対訳	95k	22.03	8.89
BCCWJ 精選あり + Tatoeba 対訳	95k	23.24	11.43
Tatoeba 精選なし	95k	24.65	13.67
Tatoeba 精選あり	95k	25.15	13.73
Tatoeba 精選なし + Tatoeba 対訳	95k	27.87	9.78
Tatoeba 精選あり + Tatoeba 対訳	95k	**28.77**	**15.80**
Tatoeba 対訳 (ベースライン)	10k	19.10	9.75
Tatoeba 精選なし	10k	14.66	4.11
Tatoeba 精選あり	10k	17.19	7.90
Tatoeba 精選なし	50k	21.73	10.08
Tatoeba 精選あり	50k	23.43	13.44
Tatoeba 精選なし + Tatoeba 対訳	50k	27.55	11.37
Tatoeba 精選あり + Tatoeba 対訳	50k	28.64	14.03

そこで，単言語コーパスから作られた擬似対訳コーパスの質と量が機械翻訳の結果にどの程度で影響を与えるかを調べるために BCCWJ の 2,355,503 文の内 10k，50k，95k、100k，500k，1M，2M の文を学習データとして抽出し，それぞれの PBSMT と NMT の翻訳精度について調べた．

4.3 実験結果

表 2 にロシア語-日本語言語対で BCCWJ から作られた擬似対訳コーパスのみを用いた機械翻訳の結果と，擬似対訳コーパスに Tatoeba のロシア語-日本語対訳コーパスを加えたものを用いた機械翻訳の結果を示す．疑似対訳コーパスのみを用いて機械翻訳を行っても，文数を増やすにつれ，BLEU が上がることがわかる．さらに，疑似対訳コーパスを精選することで翻訳精度が上がることが示された．具体的には，擬似対訳コーパスの 581,401 文が BLEU+1>0 であり，500k 文までで学習した際に，PBSMT で+3 ポイント，NMT で+4 ポイント上がった．一方で，BLEU+1 が 0.00 になっている文対が含まれている 1M，2M 文対では BLEU が下がっている．

複数の先行研究で翻訳精度を上げるために対訳コーパスを加える手法がある [10]，[6]．同様に，擬似対訳コーパスに対訳コーパスを加えると翻訳精度がどの程度で上がるかを実験した．全ての実験結果において，擬似対訳コーパスのみを用いた機械翻訳の BLEU より，擬似対訳コーパスに 10k 文対の対訳コーパスが含まれているコーパスを用いた機械翻訳の BLEU が高い，PBSMT で+10 ポイントまで，NMT で+5 ポイントまで上がった．また，いずれの条件においても，PBSMT は NMT より BLEU が高くなっている．

表 3 にベースラインとしてピボット機械翻訳と 10k 文対で学習されたロシア語-日本語の直接翻訳の結果を示す．ベースラインと比較するために，ピボット機械翻訳に用いられた文数に合わせて提案手法の実験結果を示す．PBSMT では，異なるドメインの BCCWJ の単言語コーパスから作られた疑似対訳コーパスを用いた実験結果がピボット機械翻訳の BLEU を上回っている．さらに，擬似対訳コーパスの精選を行ったところ，翻訳精度が上がっている．同じ Tatoeba ドメインで実験の単言語コーパスから作られた疑似対訳コーパスを用いた際には，BLEU がピボット翻訳の 2 倍になっている．さらに，同じドメインの同じ 10k の精選した擬似対訳コーパス (BLEU+1>0.56) では，翻訳精度がベースラインの対訳コーパスでの翻訳精度との差は 1.91 ポイントとなる．単言語コーパスの文数を 50k まで増やすと，58,528 文目以降は BLEU+1 が 0.00 となるため，翻訳精度はベースラインを+4.33 ポイントまで上回る．表 4 より，文精選によって流暢な出力が得られたことがわかる．

表 4: 対訳コーパスの分量を揃えて 95k 文対で学習したモデルの出力例

ソース: Билл мой самый близкий друг .		正解: ビル は 私 の 一番 の 親友 です 。
モデル	PBSMT による出力	NMT による出力
Tatoeba 対訳 (ベースライン) 10k	ビル は 私 の 中 で 一番 близкий 友達 です 。	ビル は 私 の 一番 背 の 背 が 大好き です 。
Tatoeba ピボット (ベースライン)	私 は その Билл близкий 友達 です 。	私 は その Билл близкий 友達 です 。
BCCWJ 精選なし	ビル は 私 の 一番 近く の 友人 。	私 の <unk> は 私 の 子供 だ 。
BCCWJ 精選あり	ビル は 私 の 最大 の 親友 で ある 。	<unk> は 私 の 好き な もの です 。
BCCWJ 精選なし + Tatoeba 対訳	ビル は 私 の 一番 という 友達 です 。	ビル は 彼女 が 一番 一番 人 。
BCCWJ 精選あり + Tatoeba 対訳	ビル は 私 の 最大 の 親友 で ある 。	ビル は 私 の 友達 が 速い 。
Tatoeba 精選なし	ビル は 私 の 最も 親しい 友人 です 。	ビル は 私 の 一番 背 の 高い 。
Tatoeba 精選あり	ビル は 私 の 最も 親しい 友人 です 。	ビル は 私 の 親しい 友人 です 。
Tatoeba 精選なし + Tatoeba 対訳	ビル は 私 の 最も 親しい 友人 です 。	トム は 私 の 友達 の 人 です 。
Tatoeba 精選あり + Tatoeba 対訳	ビル は 私 の 最も 親しい 友人 です 。	ビル は 私 の 一番 親しい 友人 です 。

5 考察

表 2 の実験条件では，1M-2M 文対で学習された翻訳精度は少ない文対で学習した際の翻訳精度より低い．理由は，疑似対訳コーパスに BLEU+1 が 0.00 であるようなノイズが含まれているためであると考えられる．精選された擬似対訳コーパスのうち，500k 文対は BLEU+1 が 0 より大きく，それらの文対で学習すると高い翻訳精度が得られるが，後でノイズが含まれるため翻訳精度が下がると考えられる．精選されていないランダムな疑似対訳コーパスにおいても，規模が大きくなるにつれノイズの量も増えるので，結果的に翻訳精度が下がると考えられる．

NMT の BLEU は PBSMT と比較して，大きく下回った．これは，NMT が，学習に大規模コーパスを要するため，もしくは質の十分でない文対を多く含みうるコーパスでは，PBSMT に比べ学習が困難であるからではないかと考えられる．また，目的言語側の単言語コーパスを原言語に翻訳する際に PBSMT である Google Translate を用いて機械翻訳したため，NMT の翻訳精度が PBSMT の翻訳精度より低くなった可能性がある．

6 おわりに

対訳コーパスを持っていなくても，擬似対訳コーパスを作成することで翻訳モデルを学習可能であることが示された．精選されたコーパスがランダムな対訳コーパスより翻訳精度が高い結果が得られる．このことから，翻訳精度がデータの量だけではなく，データの質にも大きく依存することが示された．

今後は，PBSMT と比較して NMT の精度が低い問題が擬似対訳コーパスの量，擬似対訳コーパス内のノイズの量，擬似対訳コーパスを生成する機械翻訳モデルによる影響なのか，それともその他の原因なのかを明確にする必要がある．

参考文献

[1] Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. Improving statistical machine translation by paraphrasing the training data. In *IWSLT*, 2008.

[2] Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*, 2016.

[3] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ACL*, pages 728–735, 2007.

[4] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *ACL-IJCNLP*, pages 1723–1732, 2015.

[5] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL-HLT*, pages 866–875, 2016.

[6] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. *arXiv preprint arXiv:1606.04164*, 2016.

[7] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

[8] Chin-Yew Lin and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING*, pages 501–507, 2004.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96, 2016.

[11] Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*, pages 1535–1545, 2016.

[12] Barret Zoph and Kevin Knight. Multi-source neural translation. In *NAACL-HLT*, pages 30–34, 2016.

[13] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, pages 1568–1575, 2016.