双方向翻訳モデルの相互学習による対訳語彙の教師なし獲得過程の調査

谷川 琢磨　　秋葉 友良　　塚田 元
豊橋技術科学大学
{tanigawa.takuma.fu, akiba.tomoyoshi.tk, tsukada.hajime.hl}@tut.jp

概要

本稿では，データ拡張手法である Iterative Back-translation（IBT）を用いたドメイン適応による単言語資源からの知識獲得について調査を行った．我々の先行研究では，2 言語の単言語コーパスに分かれて出現する互いに翻訳関係にある語のペア (対訳語彙) であっても，IBT を繰り返すことで次第に対訳として翻訳できるようになることを明らかにした．今回の実験では，それらの対訳語彙について，詳しい獲得の過程や条件について掘り下げを行った．その結果，対訳語彙の獲得には，単言語データでの対訳語彙の出現回数が関係していることが明らかになるとともに，対訳語彙が出現する文中の文脈が関係していることが示唆された．

1　はじめに

ニューラル機械翻訳（NMT）の翻訳精度には学習データに使用する対訳コーパスの量と質が大きく関わっている．しかし，特定のドメインでは十分な量の対訳コーパスを用意することが困難であるという問題がある．そこで，比較的収集が容易な単言語コーパスを用いる手法が提案されている．そのような手法の一つとしてデータ拡張手法である Iterative Back-translation(IBT)[1][2][3][4] を用いたドメイン適応が知られている．IBT は，翻訳対象言語対の 2 つの単言語コーパスを相互に逆翻訳とモデルの更新を繰り返し行うことで，疑似対訳データと翻訳モデルの質を向上させることができる．

我々の先行研究 [5] では IBT による単言語コーパスからの知識獲得の過程を，2 言語の単言語コーパスのみに出現する互いに翻訳関係にある語のペア (対訳語彙) の獲得を調べることによって調査した．その結果，対訳語彙は IBT の反復を繰り返すごとに獲得していき，最終的には獲得可能な 6 割以上の対訳語彙を獲得できていることがわかった．しかし，対訳語彙の詳しい獲得過程や獲得条件などについては明らかにはできていなかった．そのため本稿では，対訳語彙の獲得についてさらなる掘り下げを行った．具体的には，単言語データでの対訳語彙の出現頻度が対訳語彙獲得率に与える影響，対訳語彙ごとの IBT の逆翻訳結果上での獲得率，獲得できなかった対訳語彙について調査を行った．その結果，対訳語彙の獲得過程については，テストデータで獲得できていた対訳語彙の約 7 割が，IBT の 1 回目の逆翻訳時点でどちらかの翻訳方向で少なくとも 1 箇所で翻訳結果に出力できている事がわかった．また，単言語コーパスのどこかで 1 度でも翻訳に成功すれば，その後の反復で次第に他の箇所でも翻訳が成功するようになり，対訳語彙獲得につながることが観察された．さらに，獲得の条件については，単言語データ上での出現回数が重要であることがわかった．加えて，対訳語彙の獲得には語の出現する文脈が大きく関係していることも示唆された．

2　Iterative Back-translation（IBT）

Iterative Back-translation(IBT) によるドメイン適応手法の手順を説明する．ここで，X と Y はそれぞれの言語を示し，言語 X から Y の翻訳を X-Y，Y から X への翻訳を X-Y と記す．

1. ドメイン外の対訳コーパス C_X^{out} と C_Y^{out} を用いて，$Model_{X-Y}0$ と $Model_{Y-X}0$ を学習する．
2. i を初期化して以下を反復する．
   2.1 ドメイン内単言語コーパス C_Y^{in} を $Model_{Y-X}i$ により翻訳し，疑似対訳コーパス（$C_X'^{in}, C_Y^{in}$）を作成する．疑似対訳コーパスと（C_X^{out}, C_Y^{out}）を結合した学習データを用いて，$Model_{X-Y}i$ から Fine-tuning を行い $Model_{X-Y}(i+1)$ を学習する．
   2.2 ドメイン内単言語コーパス C_X^{in} を

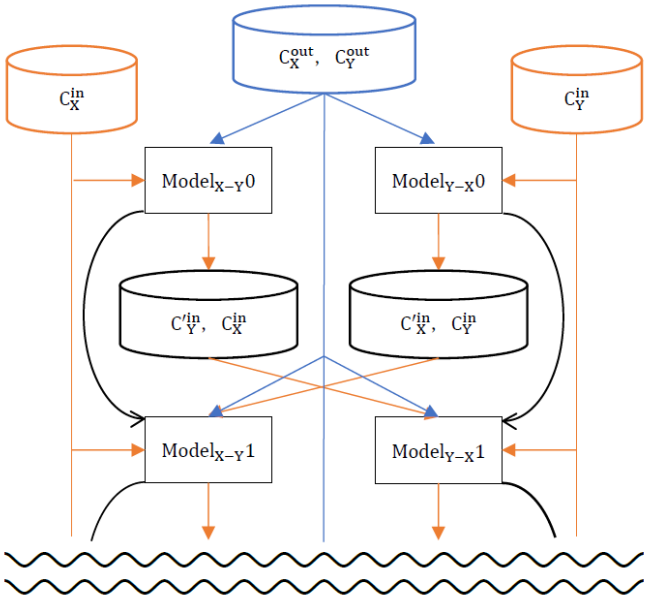

図 1　Iterative Back-translation の手順

$Model_{X-Y}$i により翻訳し，疑似対訳コーパス（C'^{in}_Y, C^{in}_X）を作成する．疑似対訳コーパスと（C^{out}_Y, C^{out}_X）を結合した学習データを用いて，$Model_{Y-X}$i から Fine-tuning を行い $Model_{Y-X}$(i + 1) を学習する．

2.3 i ← i + 1

IBT により対訳語彙が獲得される原理は，次のように説明できる．言語 X の語 x と Y の語 y が対訳語彙 (x, y) であったとする．IBT の初期の翻訳モデル $Model_{X-Y}0$ では X の単言語コーパスに出現する x を y に翻訳する事はできないはずである．しかし，逆翻訳によって生成される擬似対訳には，X 言語側には単言語コーパスの x を含む文そのままが，Y 言語側には (y は出現しないとしても)“x のコンテキストの翻訳”(≈“y のコンテキスト”) を含む文が得られる．この擬似対訳を使って $Model_{Y-X}1$ を学習すると，y および “y のコンテキスト” は，x および “x のコンテキスト” に翻訳される可能性が出てくる．IBT の反復を繰り返すにつれて，次第に (x, y) を含む擬似対訳が生成される確率も上昇し，最終的に対訳語彙が獲得されると考えられる．

3　調査手法

ドメイン内知識の獲得の過程を確認するために，本実験では対訳語彙の獲得を調査した．言語 X と言語 Y の対訳語彙獲得を調査するために行った手順は以下のとおりである．

表 1　種類ごとの対訳語彙の例

種類	英語	日本語
自明	MIC	MIC
カタカナ	intranet	イントラネット
漢字	transfusion	輸血
その他	convulsion	けいれん

1. ドメイン内の学習データに存在する単語とドメイン外の学習データに存在する単語について，それぞれのドメインの学習データから単語を列挙することで調べる．それによりテストデータ中のドメイン内にのみ存在する単語を特定し，そのようにして特定した単語の集合を D_X と D_Y とする．
2. 単語アライメントツール (Moses[6] 付属の GIZA++) を用いて言語 X と言語 Y のテストデータ間の単語アライメントを作成する．そのアライメントの単語対応がともに D_X，D_Y に含まれているもの $T = \{(w_X, w_Y) | w_X \in D_X, w_Y \in D_Y\}$ を対訳語彙とする．
3. IBT の翻訳モデルによってテストデータの翻訳を行い，T の対訳語彙の入力側の単語 $w_X \in D_X$ それぞれに対して，翻訳結果に対応した単語 $w_Y \in D_Y$ が出力されていれば，その対訳語彙が獲得できているとする．
4. モデルごとの対訳語彙の獲得を比較するために，全テストデータの対訳語彙の入力延べ数に対する対訳語彙を獲得できた割合を対訳語彙獲得率と定義する．

4　実験

先行研究の BPE での実験結果をもとにさらなる対訳語彙の調査を行った．まず，対訳語彙について幾つかの種類に区別を行い，それぞれの種類ごとの対訳語彙の獲得について調査を行った．区別した種類として，対訳語彙のペアがそれぞれ同じ単語である「自明」，対訳語彙のペアが自明な語以外で翻字の関係にある「カタカナ」，日本語側の対訳語彙がすべて漢字である「漢字」，以上のどれにも当てはまらない「その他」の 4 種類に区別した．種類ごとの対訳語彙の例を表 1 に示す．以降の分析は，「その他」を除く前者 3 種類について行った．

表 2　モデルごとの逆翻訳できた対訳語彙の割合の一例

ソース側対訳語彙	ターゲット側対訳語彙	0	1	2	3	4	5	6	7	8
amplifier	増幅器	0.00	0.00	0.11	0.22	0.66	0.63	0.83	0.88	0.89
増幅器	amplifier	0.00	0.07	0.13	0.52	0.63	0.82	0.84	0.87	0.87
necrotic	壊死	0.00	0.02	0.00	0.24	0.46	0.73	0.91	0.95	0.99
壊死	necrotic	0.00	0.00	0.00	0.01	0.11	0.31	0.37	0.49	0.54
coaxial	同軸	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
同軸	coaxial	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01

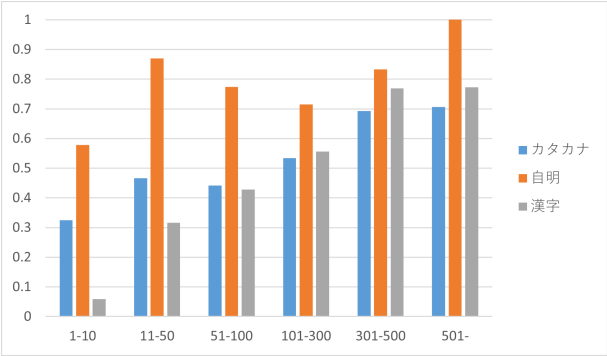

図 2　学習データでの出現回数ごとの対訳語彙獲得率

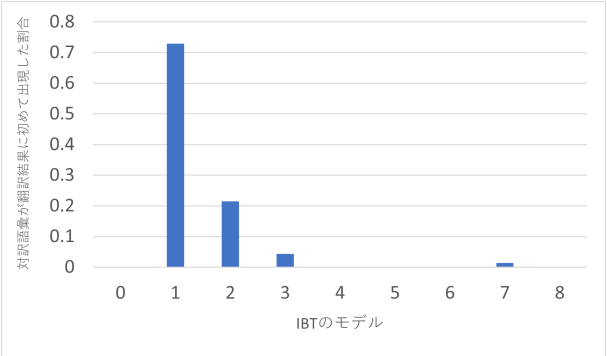

図 3　Model i ごとの初めて翻訳結果に出力できた対訳語彙の割合

表 3　誤った翻訳の一例

対訳語彙 (入力側)	対訳語彙 (出力側)	実際の翻訳結果
防臭	deodorization	antibacterial
血友病	hemophilia	friend blood
coaxial	同軸	光ファイバ
suction	吸込	噴流

4.1　学習データ中での出現回数

対訳語彙が単言語データ中に出現する回数が対訳語彙獲得率にどのように影響するかを調べるために，対訳語彙の出現回数ごとの対訳語彙獲得率調べた．英日方向でのそれぞれの種類と出現回数ごとの対訳語彙の獲得率についての結果を図 2 に示す．出現回数が多い場合「自明」が最も高く，「カタカナ」と「漢字」の対訳語彙が同じくらいの結果となった．また，「カタカナ」と「自明」は学習データ中での出現回数が少ない対訳語彙でも獲得されていた．これはこの 2 種類は対訳語彙のペアが翻字の関係であることが関係していて，自明な語の場合は特にサブワード化の影響を大きく受けているからだと考えられる．

4.2　対訳語彙の獲得過程

対訳語彙の獲得の過程について詳しく調査を行うために，単言語コーパスの逆翻訳結果を観察し，IBT の Model i がどの程度対訳語彙のソース側をターゲット側に翻訳できているかの割合 (翻訳率) を調査した．調査の対象とする対訳語彙については，サブワードの影響を受けづらい「漢字」について，人手でチェックを行ったもののみを対象とした．獲得過程の結果の一例を表 2 に示す．

対訳語彙の「増幅器」と「amplifier」の例から見て取れるように，翻訳率は IBT を繰り返すごとに上昇していた．そのため，対訳語彙はある時点で急に獲得されるというわけではなく，反復を繰り返すことによって徐々に獲得されていくということが分かる．一方，対訳語彙の「同軸」と「coaxial」の場合は，Model8 までほとんど割合が上昇することはなかった．実際，この対訳語彙はテストデータ上でも獲得に失敗している．

英日，日英の両方向でのテストデータで獲得することができていた対訳語彙について，それが始めて逆翻訳結果に出現した Model i を調べた．結果を図 3 に示す．この結果から，約 7 割の対訳語彙は Model 1 から疑似対訳コーパス上に出現しているなど，多くの対訳語彙は早い段階から獲得が進んでい

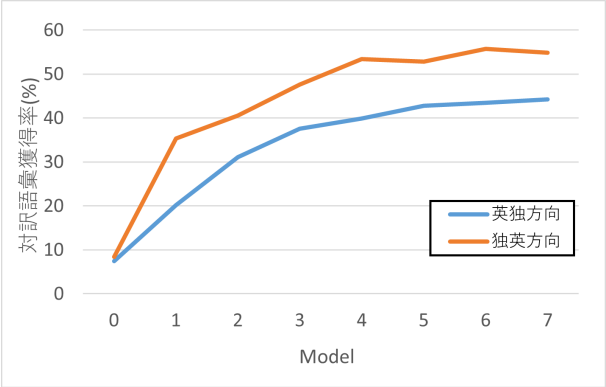

図 4　英独実験での対訳語彙獲得率

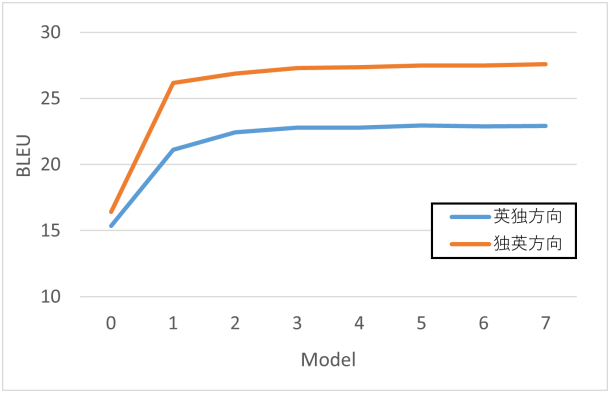

図 5　英独実験での BLEU

表 4　英独実験で獲得できた対訳語彙の例

英語	ドイツ語
codecision	Mitentscheidung
Ombudsman	Bürgerbeauftragte
Interinstitutional	Interinstitutionelle
audiovisual	audiovisuellen
KFOR	KFOR
Lulling	Lulling

ることが分かる．

4.3　獲得できなかった対訳語彙の調査

獲得できていない対訳語彙について注目して調査を行った．具体的には，獲得できなかった対訳語彙が本来翻訳される単語の代わりに何を翻訳しているのかについて調べた．結果の一例を表 3 に示す．

ほとんどの獲得できなかった対訳語彙において，本来翻訳されるべき正しい単語の代わりに違う単語が学習されていた．表の対訳語彙の「coaxial」は「同軸」と翻訳される代わりに「光ファイバ」と翻訳されることが多かった．これは「光ファイバ」と「同軸」が出現するような文の文脈が近く，なおかつ「光ファイバ」のほうが学習データ中での出現頻度が高いため，間違えて学習されてしまったと考えられる．これらの調査結果から，語の出現する文中の文脈が対訳語彙の獲得に大きく影響していることが示唆された．

一方，対訳語彙の「血友病」の場合は，「血友」をサブワード分割した「血」と「友」の訳語である「blood」と「friend」に翻訳されていた．本来の対訳「hemophilia」とも他の語とも文脈に基づく結び付けに失敗したと考えられる．

4.4　英独翻訳での対訳語彙獲得の調査

異なる言語間での対訳語彙の獲得を確認するために，英語とドイツ語間での IBT によるドメイン適応実験を行った．データセットには WMT14 データセットを使用して，News Commentary コーパスから Europarl コーパスへのドメイン適応実験を行った．Europarl コーパスは単言語コーパスとして使用するために，全文を半分に分割して英語には前半部分を使用，ドイツ語には後半部分を使用した．テストデータには WMT06，07，08 の Europarl の test データを合わせた対訳 6000 文を使用した．実験条件については先行研究と同様して行った．対訳語彙獲得率の結果を図 4，BLEU を図 5 に示す．獲得された対訳語彙を表 4 に示す．実験結果より，日本語英語間での実験と同様に対訳語彙を獲得できるということが確認できた．

5　結論

本稿では，IBT による対訳語彙の獲得の過程と条件について詳しく調査を行った．IBT の対訳語彙の獲得の過程については，IBT の反復を重ねるごとに徐々に逆翻訳結果に対訳語彙が出現するようになり，その疑似対訳コーパスを学習することで獲得していることが分かった．獲得条件については，対訳語彙が単言語コーパス上に出現する回数が，特に漢字の対訳語彙の場合は多いほど獲得には有利であることが示された．また，対訳語彙の獲得には文脈が大きく関係しているということが示唆された．

謝辞

本研究は JSPS 科研費 18H01062 の助成を受けた.

参考文献

[1] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In **Proceedings of the 2nd workshop on neural machine translation and generation**, pp. 18–24, 2018.

[2] Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data. In **Proceedings of the AAAI Conference on Artificial Intelligence**, pp. 555–562, 2018.

[3] 森田知熙, 秋葉友良, 塚田元. 双方向の逆翻訳を利用したニューラル機械翻訳の教師なし適応の検討. 情報処理学会研究報告 2018-NL-238 (第 5 回自然言語処理シンポジウム), pp. 1–5, 2018.

[4] 森田知熙, 秋葉友良, 塚田元. ニューラル機械翻訳の反復的逆翻訳に基づくデータ拡張のための混成サンプリング手法. 電子情報通信学会論文誌, Vol. J106-D, No. 04, Apr. 2023. (to appear).

[5] 谷川琢磨, 秋葉友良, 塚田元. Iterative Back-translation は対訳語彙を獲得できるか? 言語処理学会, 第 28 回年次大会, pp. 354–359, 2022.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, 2007.