

技術翻訳のための 機械翻訳のカスタマイズ手法

2023年2月20日
(株) クロスランゲージ
開発部
高橋 博之

カスタマイズの必要性

近年、NMT（ニューラル機械翻訳）により機械翻訳の訳質が飛躍的に向上している。

しかし、技術翻訳に使用するにはまだ問題がある。

- ✕ 専門用語を正しく訳せない。
- ✕ 文体や言い回しが不自然。



翻訳したい文書に合わせた機械翻訳エンジンのカスタマイズが必要

カスタマイズの手法

- 辞書登録

- ⚠ 名詞のみ

- ⚠ 常にその訳が使用される（文脈による訳し分けができない）
（人名、商品名などの固有名詞には有効な方法）

- 対訳文学習

- （NMTは対訳文からの学習で「**翻訳モデル**」を作り、それを用いて翻訳）
翻訳したい分野の対訳文を用意して追加学習させることで翻訳モデルを強化

- ✓ 単語の訳だけでなく、文体や言い回しも学習

- ✓ 文脈に応じた訳し分けが可能

- ⚠ 学習用の対訳文が必要

- NMTは数千万～数億文レベルの対訳文学習で作られるが、
カスタマイズには**数万文程度で十分**



「アダプテーション」
「ファインチューニング」
とも呼ばれる技術

対訳文学習によるカスタマイズの実証実験

以下のサービスについて調査

- Google AutoML Translation
- Microsoft Custom Translator

両社とも一般向け機械翻訳サービスを提供している。

一般向け機械翻訳をベースに対訳文学習によるカスタマイズが可能。

クラウドコンピューティングサービス上で実行。

(Google Cloud Platform/Microsoft Azure)

カスタマイズ手順

両社ともクラウド上で作業可能でWeb ベースのUIがある。



Step1 学習用対訳文用意

- 1文ずつの対訳にする
- テキスト形式への変換



Step4 翻訳モデル学習

- 学習は完全に自動化されている



Step2 クラウド環境設定

- アカウント作成等



Step5 自動評価(BLEU)

- 学習前、学習後の翻訳結果のBLEU値が算出される。



Step3 対訳文をサーバにアップロード



Step6 翻訳

- 学習後のカスタム翻訳モデルをサーバにデプロイして、APIを使って翻訳

画面イメージ(MS)：データセット

Cognitive Services | Language Studio | Custom Translator Preview

<<

Workspaces

cross-dev

med-ej-30k

Manage documents

Train model

Model details

Test model

Publish model

Workspace settings

Manage documents

Documents are stored inside your workspace and are used when training models. All documents in the relevant languages are automatically available for use within a project. [Learn more about documents](#)

+ Add document set ↓ Download 🗑 Delete

Name	Created on	Type	English sentences	Japanese sentences	Created by
long_train_30k_je	06/09/2022	Training	30,000	30,000	dev cross
long_test_je	06/09/2022	Testing	2,000	2,000	dev cross
long_dev_je	06/09/2022	Tuning	2,000	2,000	dev cross
long_train_30k_ej	06/10/2022	Training	30,000	30,000	dev cross
long_test_ej	06/10/2022	Testing	2,000	2,000	dev cross
long_dev_ej	06/10/2022	Tuning	2,000	2,000	dev cross

画面イメージ(MS)：学習（トレーニング）実行

Cognitive Services | Language Studio | Custom Translator Preview

<<

Workspaces

cross-dev

med-ej-30k

Manage documents

Train model

Model details

Test model

Publish model

Workspace settings

Model details > med_30k_ej

View test Unpublish Update regions Duplicate model

Training succeeded

PUBLISHED

BLEU score

▲ 44.29
(14.84)

Baseline BLEU: 29.45

Training succeeded on: 6/10/2022 13:07PM

Total training time: 4 hour(s) 56 minute(s)

Training: 29,971 Tuning: 1,998 Testing: 2,000 Dictionary: 0

System generated test set: Yes ⓘ

Category ID:

[Learn how to evaluate a model's BLEU score](#)

Document sets used for training

Name	Created on	Type	English sentences	Japanese sentences	Aligned sentences	Used sentences
long_train_30k_ej	06/10/2022	Training	30,000	30,000	30,000	29,971
long_test_ej	06/10/2022	Testing	2,000	2,000	2,000	2,000
long_dev_ej	06/10/2022	Tuning	2,000	2,000	2,000	1,998

画面イメージ(MS)：翻訳結果

Cognitive Services | Language Studio | Custom Translator Preview

Q

<<

Workspaces

cross-dev

med-ej-30k

Manage documents

Train model

Model details

Test model

Publish model

Workspace settings

Test model > med_30k_ej

Download results

Showing 1-20 of 2000 results

	English	Japanese
1	Magnetic resonance imaging (MRI) scans of the thorax and upper abdomen do not appear to yield advantages over CT scans.	Reference: 胸部および上腹部の磁気共鳴画像法（MRI）スキャンは、CTスキャンよりも優れていないようである。 New model: 胸郭および腹部上部の磁気共鳴画像法(MRI)スキャンは、CTスキャンよりも優れているとは思われない。 Baseline model: 胸郭と上腹部の磁気共鳴画像(MRI)スキャンは、CTスキャンよりも優位性を生み出さないようです。
2	Be aware that marrow plasma cell distribution may vary in different sites.	Reference: 骨髓形質細胞の分布は、別の部位においては異なる場合があることに注意する。 New model: 骨髓形質細胞の分布は部位によって異なる可能性があることに注意すること。 Baseline model: 骨髓の血漿細胞の分布は、異なる部位で異なる可能性があることに注意してください。
3	Of 1057 patients enrolled, 298 had a history of sP_ carinii pneumonia.	Reference: 登録した患者 1,057 人中 298 人にニューモシスチスカリニ肺炎の既往を認めた。 New model: 登録した患者1,057人中、298人がカリニ肺炎ofsP_病歴があった。 Baseline model: 登録された1057人の患者のうち、298人がカリニ肺炎ofsP_歴史を持っていた。

対訳文数、時間、費用

- 対訳文数:
 - Google: 1000文以上(5000文以上が望ましい)
 - MS: 10,000文以上
 - 今回は **3 万文** 使用
- 時間: 3万文学習で **3～5時間** 程度
- 費用: 3万文学習で **\$70～\$130** 程度
- 学習結果（カスタム翻訳モデル）を用いた場合の翻訳料金:
 - ⚠ 通常翻訳よりも割高になる
 - Google: **\$80/100 万文字**（通常: \$20/100 万文字）
 - MS: **\$40/100 万文字**（通常: \$10/100 万文字）

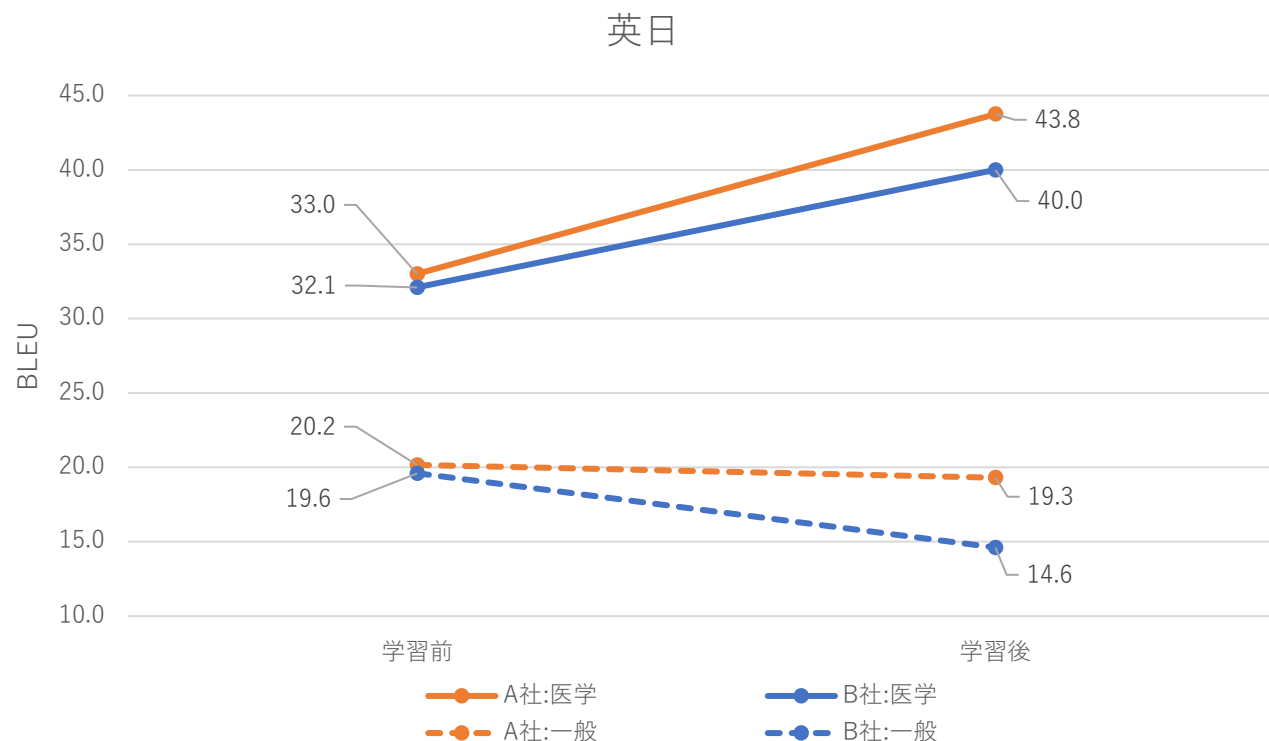
※2022年6月時点

医学論文を用いた学習（英日）

医学論文 3 万文を用いて英日・日英で学習実施
医学だけでなく一般分野の訳質の変化も見る

英日	評価文	学習前	学習後
A社	医学	33.0	43.8
	一般文	20.2	19.3
B社	医学	32.1	40.0
	一般文	19.6	14.6

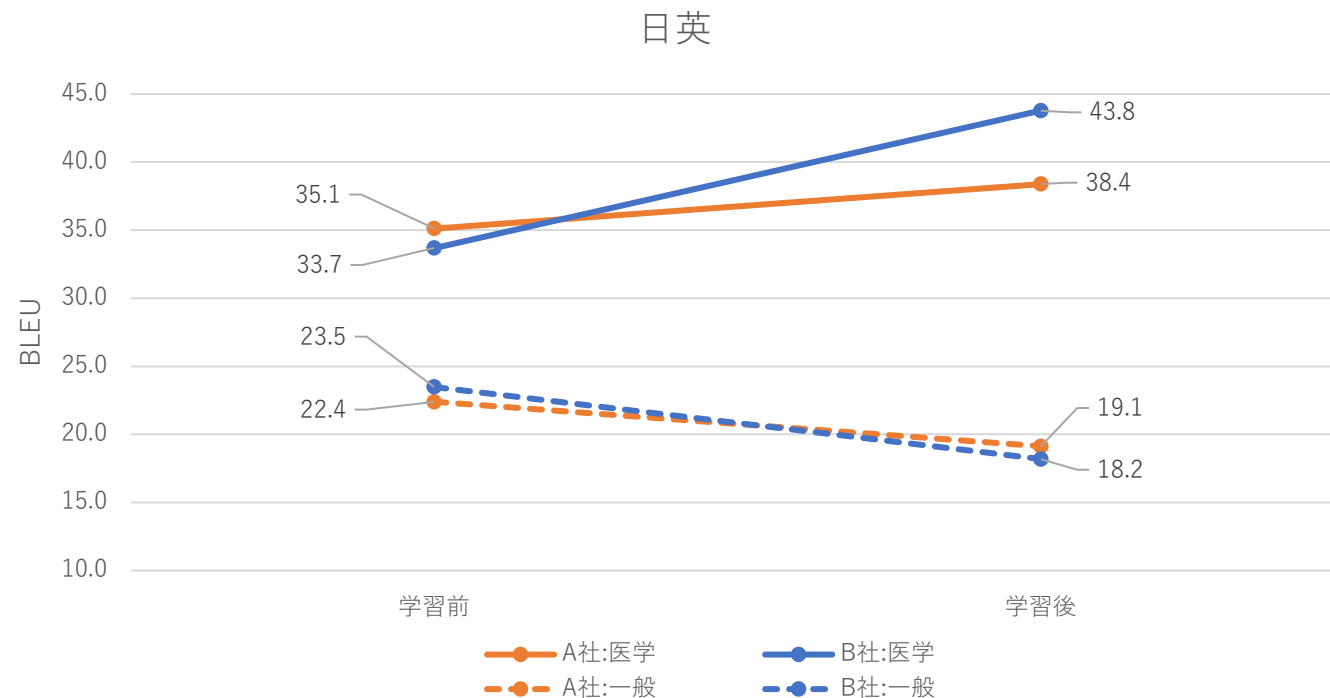
（数値は BLEU）



医学論文を用いた学習（日英）

日英	評価文	学習前	学習後
A社	医学	35.11	38.38
	一般文	22.39	19.12
B社	医学	33.68	43.77
	一般文	23.48	18.18

（数値は BLEU）



BLEU

- BLEU は機械翻訳結果の自動評価手法の一つ
- 参照訳（人間が作った正しい訳）と機械翻訳結果がどのくらい似ているかで評価する。

BLEU スコア	解釈
0～10	ほとんど役に立たない
10～20	主旨を理解するのが困難である
20～30	主旨は明白であるが、文法上の重大なエラーがある
30～40	理解できる、適度な品質の翻訳
40～50	高品質な翻訳
50～60	非常に高品質で、適切かつ流暢な翻訳
60～100	人が翻訳した場合よりも高品質であることが多い



カスタマイズにより
BLEU30台（理解できる）
から40台（高品質）に！

（ <https://cloud.google.com/translate/automl/docs/evaluate?hl=ja> より）

改善例

医学用語が正しく訳せるようになる。

原文	IV期の口唇癌および口腔癌
学習前	Stage IV lip and oral cancer ⚠
学習後	Stage IV Lip and Oral Cavity Cancer ✓

医学論文では患者を「人」ではなく「例」で数える。

原文	Thirty patients (81%) had healthcare-associated infections; 18 (60%) of these 30 were infected with isolates carrying markers of CA-MRSA strains.
学習前	30人の患者 (81%) が医療関連の感染症を患っていました。これら30のうち18 (60%) は、CA-MRSA株のマーカを保有する分離株に感染していた。⚠
学習後	患者30例 (81%) が医療関連感染症であった；この30例中18例 (60%) が、CA-MRSA株のマーカを保有する分離株に感染していた。✓

悪化例

• 分野外の文の訳質は悪化することがある

用語選択がその分野寄りになる

原文	To ensure that the noodles will go perfectly with the soup, their thickness and length are carefully calibrated .
学習前	麺がスープと完璧に合うように、その太さと長さは慎重に 調整 されています。
学習後	スープの厚みと長さを注意深く キャリブレーション して、完璧なスープを作り上げます。

対象分野外の表現が訳せなくなる

原文	She's been putting in more hours at the office recently.
学習前	彼女は最近、オフィスでより 多くの時間を費やして います。
学習後	彼女は最近オフィスに もっと時間を置いて いる。

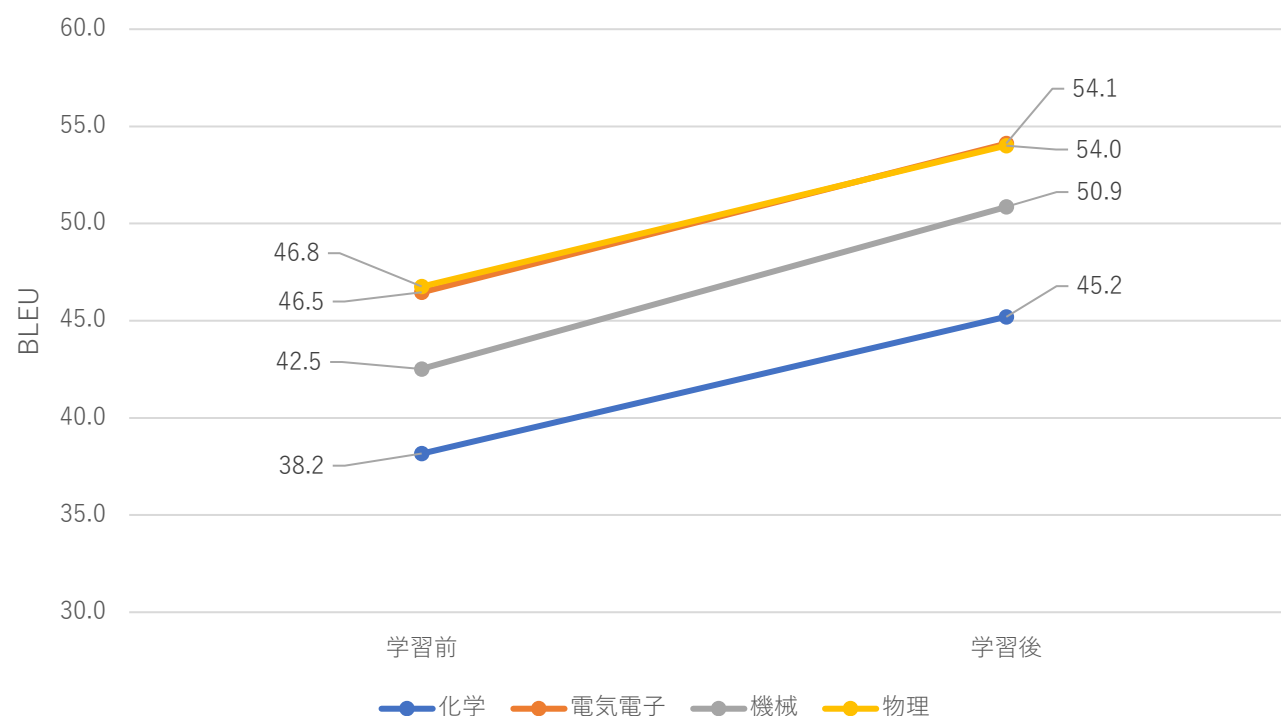
医学以外の分野での実証

- 各種の専門的な文のソースとして特許公報を利用（日→中翻訳）
- 化学、電気電子、機械、物理の4分野でそれぞれ10万文学習

分野	学習前	学習後
化学	38.2	45.2
電気電子	46.5	54.1
機械	42.5	50.9
物理	46.8	54.0

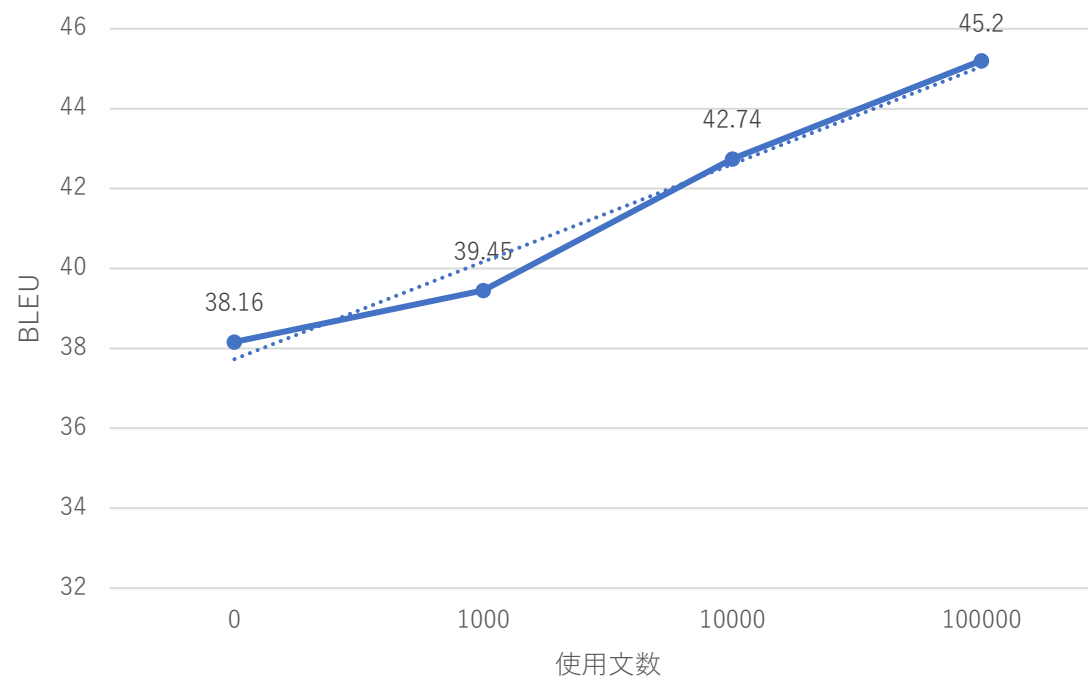
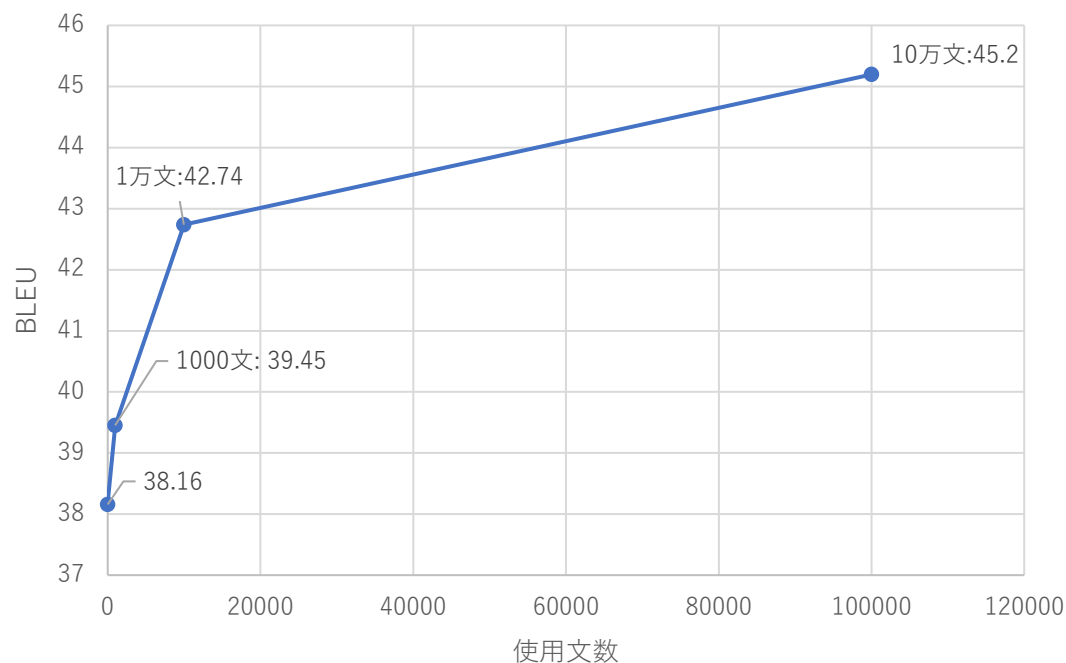
（数値は BLEU）

各分野とも +7~8 程度の向上



学習に使用する対訳文の数と訳質の関係

- 化学分野で学習に使用する文数を1000、1万、10万と変化させた



対訳文が多いほど訳質は向上するが、伸びは鈍化する。

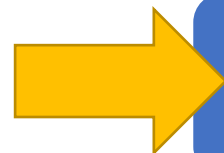
適切な分野の学習文を使うことの効果

分野分けされていない文で学習するとどうなるか？

特許文 10 万文で学習した一つのモデルで 4 分野のテスト文を翻訳

分野	学習前	学習後 (分野分けなし学習)	学習後 (分野別学習)
化学	38.2	43.8	45.2
電気電子	46.5	53.4	54.1
機械	42.5	49.8	50.9
物理	46.8	53.4	54.0

分野別に学習すると+7~8 程度向上するのに比べ、
分野分け無しだと+6~7程度の向上に留まる



学習用データは翻訳対象と同じ分野の文を使うとより効果的

まとめ

- NMT は対訳文を用いた学習によりカスタマイズできる。
 - 翻訳したい文と同分野の学習文を用意する必要がある。
 - 1 万文程度の学習文で一定の効果が得られる。
 - 学習後はその分野に特化した翻訳モデルとなり、対象以外の分野の訳質は悪化することがある。