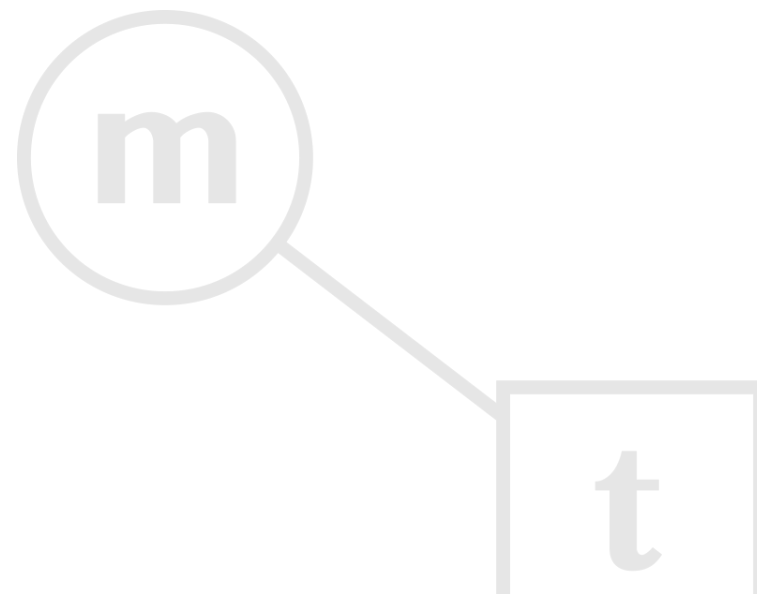




翻訳Tips

株式会社みらい翻訳
Mirai Translate, Inc.

〒150-0002 東京都渋谷区渋谷 2-22-3 渋谷東口ビル2F
TEL: 03-6434-1340 FAX: 03-6434-1590
<https://miraitranslate.com>



0. [はじめに](#)

1. [ニューラル機械翻訳の性質](#)

2. [翻訳しやすい原文をつくるために](#)

3. [テキスト翻訳とファイル翻訳](#)

4. [ユーザ辞書](#)

参考：[翻訳までの流れ](#)

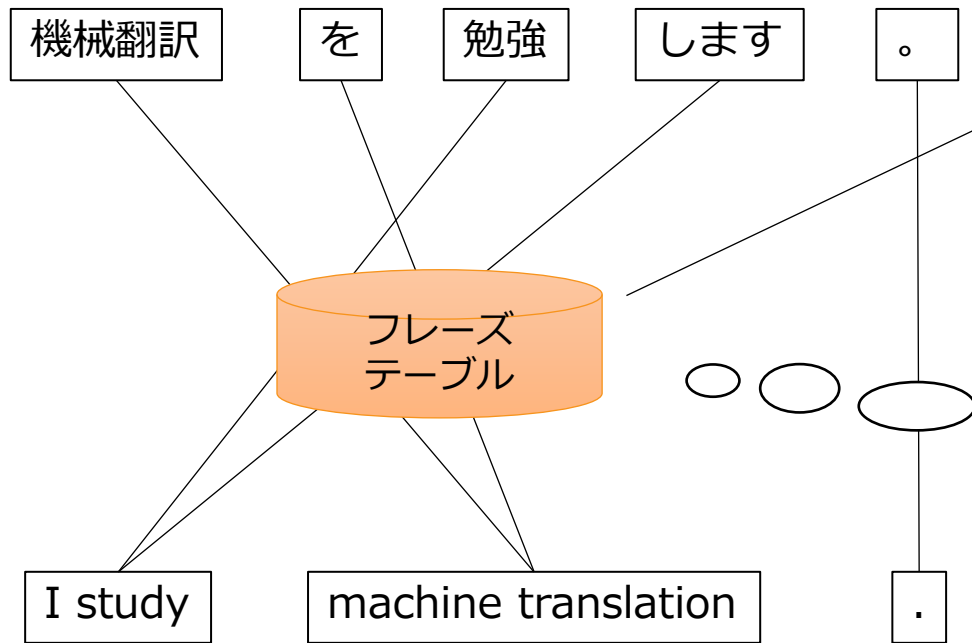
*) 本資料の内容は、今後の機械翻訳エンジンのアップデートによって変わる可能性があります。

0. はじめに (今までの機械翻訳の概要)

2010年代前半までは「機械翻訳」といえばフレーズベース統計的機械翻訳 (Phrase-Based Statistical Machine Translation, PBSMT) と呼ばれる、単語列単位の組み替えで翻訳文生成する技術が主流でしたが、単語列単位で切り貼りして文を構成するため、全体で見ると流暢性に欠けるという問題点がありました。

従来技術 (PBSMT) のベースであるSMTによる翻訳

注意：説明のために大幅に簡略化しています



コーパス：データベース化された大規模な言語資料のことを指します
機械翻訳に用いる対訳コーパスとは様々な言語のコーパスを対応付けしたものです

en	ja
I'll stay at a hotel.	ホテルに泊まります。
A little water, please.	水を少し下さい。
:	:

対訳コーパスを学習し、最もそのように訳される確率が高いものを統計的に計算して語順を組み替えます

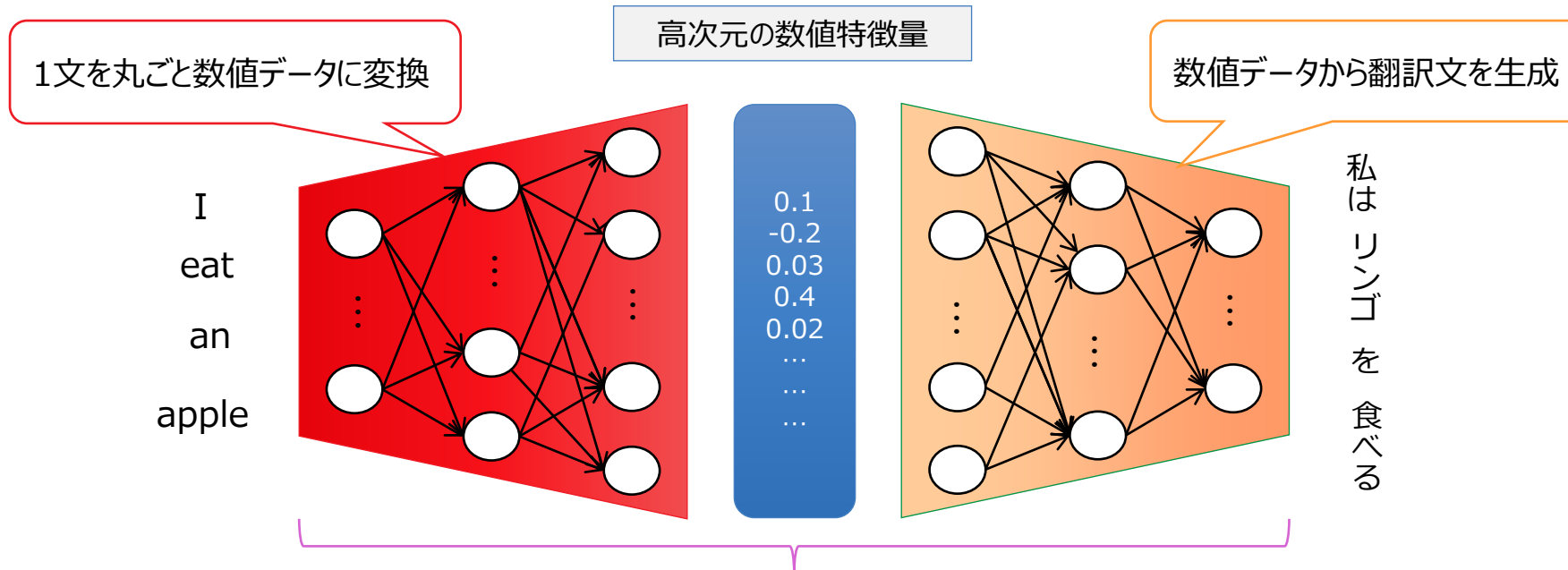
統計計算のイメージ

	studyの訳	Iと併用された時に訳される確率
1	勉強する	80%
2	研究する	●%
3	考慮	▲%

→ I study は“私は・・・勉強します”という訳になると統計的に判断

1. ニューラル機械翻訳の性質

Mirai Translatorが採用しているのはニューラル機械翻訳（Neural Machine Translation, NMT）です。NMTは単語列単位でなく、1文丸ごと数値データに変換した後、この数値データから新しい翻訳文を生成します。これによりPBSMTと比べて、より流暢な翻訳結果を生成できることが特長です。



最新の人工知能タイプの翻訳エンジンであり、原文を数値データに置き換えてから翻訳結果を生成する

1. ニューラル機械翻訳の性質 翻訳精度向上実例



従来技術であるPBSMTと比較して、NMTでは飛躍的に翻訳精度が向上しています。

原文	夕食は部屋食、朝食は食堂でという旅館もよくあります。
----	----------------------------



PBSMT 翻訳結果	Dinner is a Breakfast it is common in japanese inn that dinner at the dining room.
---------------	--

意味が分からない

初期：みらいNMTエンジン (2017) 翻訳結果	There are many Japanese-style inns in the dining room for dinner.
---------------------------------	---

原文と意味が異なっている

みらいNMTエンジン (2018) 翻訳結果	There are many ryokans that have dinner in the room and breakfast in the dining room.
------------------------------	---

Good

現在：みらいNMTエンジン (2022) 翻訳結果	There are many ryokans where dinner is served in the room and breakfast is served in the dining room.
---------------------------------	---

Very good!

1. ニューラル機械翻訳の性質 留意すべき特徴



高い翻訳精度が大きな特徴で、従来技術に比べ圧倒的に自然な文章を生成します。
一方で以下のような特徴もあるため、その性質を理解した上でご利用することをおすすめします。

留意すべき特徴

- ✓ うまくいく理由が十分に分かっていません。
→ 翻訳エラーがあった場合に修正する方法が明確ではありません。
- ✓ 原文を過不足なくカバーして翻訳しないことがあります。
→ **訳抜け**、**湧き出し**、**過剰生成**などが発生します。
- ✓ **入力の違いに敏感**で、表記揺れに弱く、句読点の有無によっても翻訳結果が大きく変わることがあります（不安定性）。
- ✓ 辞書を十分に制御して当てるのが難しい。
→ ユーザ辞書を使用する際は **4. ユーザ辞書** の項目をよく読んだ上でお使いください。

1. ニューラル機械翻訳の性質 不安定性



NMTは入力の違いに敏感で、表記揺れに弱く、句読点の有無によっても翻訳結果が大きく変わることがあります。

- 句点の有無

例) 原文	翻訳結果
Japan lowers age of adulthood for 1st time in over 140 years	日本、140年ぶりに成人年齢を引き下げ
Japan lowers age of adulthood for 1st time in over 140 years.	日本は140年以上ぶりに成人年齢を引き下げた。

- 表記揺れ

例) 原文	翻訳結果
I cannot believe you.	私はあなたを信じるできません。
I can't believe you.	信じられません。

2. 翻訳しやすい原文をつくるため



- 機械翻訳は、私たち人間のように文脈を理解することができないため、正しく翻訳するためには、シンプルで分かりやすい原文であることが必要です。
- 英語を日本語に翻訳する場合は、**曖昧な表現やイディオム、口語（話し言葉）等を避ける**ことにより、機械翻訳の精度を高めることができます。

➤ 具体的には、以下のポイントがあります。（英→日の場合）

1. 文章を正しく記載する

- スペルが正しいかを確認する
- 不要な改行は避ける

2. 大文字と小文字（特に名詞）を正しく表記する

- 大文字を正しく記載する
- 小文字を正しく記載する

3. 句読点を使用する

- 文末にピリオドまたは疑問符を付ける
- わかりやすくするためにカンマを使用する

4. 曖昧な表現を避ける

- 曖昧な表現を避ける
- 修飾語は適切な位置（修飾したい語の隣）に記載する

5. 文脈を考慮する必要のある表現を避ける

- 頭字語や略語を避ける
- イディオムを避ける
- 口語（話し言葉）を避ける

2. 具体例（英→日）



1. 文章を正しく記載する

- スペルが正しいかを確認する

	例) 原文	翻訳結果
Wrong	It was an amature mistake.	それは 大人の ミスでした。
Good Correct	It was an amateur mistake.	それは 素人の ミスでした。

- 不要な改行は避ける（削除してから翻訳する）

	例) 原文	翻訳結果
Wrong	We sold software, hardware, and computer parts at my former company.	ソフトウェアやハードウェアを販売し コンピュータ部品や 私の前の会社で。
Good Correct	We sold software, hardware, and computer parts at my former company.	前の会社では、ソフトウェア、ハードウェア、コンピュータの部品を販 売していました。

2. 具体例（英→日）



2. 大文字と小文字（特に名詞）を正しく表記する

- ・大文字を正しく記載する

	例) 原文	翻訳結果
Confusing	Yesterday, i visited atlas, a friend of mine who lives in yellow springs, ohio.	昨日、おひおに住んでいる友人の地図帳を見に行きました。
Good Clear	Yesterday, I visited Atlas, a friend of mine who lives in Yellow Springs, Ohio.	昨日、私はオハイオ州イエロースプリングスに住む友人のアトラスを訪ねた。

- ・小文字を正しく記載する

	例) 原文	翻訳結果
Wrong	INTEREST RATES RISE AGAIN TO COUNTER HIGHER PRICES.	金利は再び上昇し、COUNTERの物価上昇に対抗します。
Good Correct	Interest rates rise again to counter higher prices.	物価上昇に対抗して再び金利が上昇します。

2. 具体例（英→日）

3. 句読点を使用する

- 文末にピリオドまたは疑問符を付ける

	例) 原文	翻訳結果
Wrong	I wouldn't bet against her	私は彼女に賭けます
Correct	I wouldn't bet against her.	彼女に賭けるつもりはありません。

Good

Confusing	What goes up but never ever comes down	上がっても下がらない
Clear	What goes up but never ever comes down?	上がっても下がらないものは何か?

Good

- わかりやすくするためにカンマを使用する

	例) 原文	翻訳結果
Confusing	Wear sun protection gear like a hat with a wide brim and sunglasses to protect your face and eyes.	つばの広い帽子などの日焼け止めやサングラスで顔や目を保護しましょう。
Clear	Wear sun protection gear, like a hat with a wide brim and sunglasses, to protect your face and eyes.	顔や目を保護するために、つばの広い帽子やサングラスなどの日焼け止めを着用してください。

Good

2. 具体例 (英→日)

4. 曖昧な表現を避ける

- 曖昧な表現を避ける

	例) 原文	翻訳結果
Confusing	London had snow yesterday. It fell to a depth of a meter.	ロンドンでは昨日雪が降りました。1メートルの深さまで 落ちました 。
Good Clear	London had snow yesterday. Snow fell to a depth of a meter.	ロンドンでは昨日雪が降りました。 雪が 1メートルの深さまで降りました。

自然な表現ではありませんが、主語を明確にすると翻訳結果が良くなります。

- 修飾語は適切な位置 (修飾したい語の隣) に記載する

	例) 原文	翻訳結果
Wrong	The boy barely saw the black kitten with his sunglasses on .	その少年はサングラスをかけている黒い子猫を <u>かろうじて</u> 見ました。
Good Correct	With his sunglasses on , the boy barely saw the black kitten.	サングラスをかけたまま、少年はその黒い子猫を <u>かろうじて</u> 見ました。

	例) 原文	翻訳結果
Confusing	The torn student's book lay on the desk.	破れた 生徒の本が机の上に置いてありました。
Good Clear	The student's torn book lay on the desk.	その学生の 破れた 本は机の上に置いてありました。

5. 文脈を考慮する必要のある表現を避ける

- 頭字語や略語を避ける

	例) 原文	翻訳結果
Confusing	MOM sales show a marked improvement.	MOMの売上が大幅に伸びています。
Good Clear	Month-over-month sales show a marked improvement.	前月比の売上は顕著な改善を示しています。

	例) 原文	翻訳結果
Wrong	Calif. borders Ore. to the north and Nev. and Ariz. to the east.	カリフ。鉱石との境界。北にネブ、東にアリゾナ。
Good Correct	California borders Oregon to the north and Nevada and Arizona to the east.	カリフォルニア州は、北はオレゴン州、東はネバダ州とアリゾナ州に接しています。

2. 具体例（英→日）



5. 文脈を考慮する必要のある表現を避ける

- **イディオムを避ける**

	例) 原文	翻訳結果
Confusing	Cat got your tongue?	猫に舌をつかまれた?
Good Clear	Why are you not saying anything?	どうして何も言わないの?

- **口語（話し言葉）を避ける**

	例) 原文	翻訳結果
Wrong	We saved some money on the meal so we splashed out on a really nice bottle of wine.	私たちは食事代を節約したので、本当においしいワインを 一気に飲みました 。
Good Correct	We saved some money on the meal so we spent more than usual on a really nice bottle of wine.	私たちは食事代を節約したので、本当においしいワインに いつもよりたくさんお金を使いました 。

3. テキスト翻訳とファイル翻訳



翻訳を行う方法として「テキスト翻訳」と「ファイル翻訳」の2種類の方法があります。用途に応じて使い分けると便利ですが、それぞれ使用する際には注意点があります。

- 翻訳を実行する前に
 - **言語方向**が合っているか確認をしましょう。
 - **翻訳設定**を使って語調を統一すると読みやすいです（日本語への翻訳の場合）。※デフォルトだと語調は統一されません。
- 共通の注意点
 - 翻訳は1文ずつ行われますが、その際、**前後の文脈は考慮されません**（文の前後の意味を推測して訳すことは出来ません）。そのため、**1文ずつ入力してもまとめて入力しても結果は変わりません**。
- テキスト翻訳の注意点
 - 改行は文区切りとみなすため、**文の体裁を整える（レイアウト調整の）ために文の途中に改行を入れていると翻訳精度が劣化します**。（特にPDFからテキストをコピーして翻訳する場合や、メール文で所々に改行を入れていたりすると注意が必要です。）
- ファイル翻訳の注意点
 - **装飾（太字や色、ハイパーリンクなど）**があると、翻訳結果がテキスト翻訳と異なることがあります。
 - PDFファイルの翻訳では、ファイルによっては**OCRによって文字認識**をする必要があり、うまく認識できない場合翻訳精度も劣化する可能性が高くなります。PDFにする前のファイルをお持ちであれば、そちらをファイル翻訳にかける方が精度は高くなります。
 - テキスト翻訳と同様に、改行は文区切りとみなします。特にPDFファイルの翻訳では、解析時に意図していないところに改行が入る可能性があります。

3. テキスト翻訳とファイル翻訳



①言語方向を設定し、 ②翻訳設定が適切か確認してください。（日本語への翻訳の場合）

テキスト翻訳

The screenshot shows the MiraiTranslate interface. At the top, there are two sections: "テキスト翻訳" (Text Translation) and "ファイル翻訳" (File Translation). Both sections have a language direction selector (English to Japanese) and a settings icon. The text translation section shows the input "I translate from English to Japanese." and the output "私は英語から日本語に翻訳する。". The file translation section shows a dashed box with file icons and the instruction "翻訳するファイルをドラッグ&ドロップしてください".

①言語方向が合っているか

②文体をどちらにするか

The "翻訳設定" dialog box is shown. It has a close button (X) in the top right. The "表記" (Notation) section is selected, with the instruction "日本語の翻訳結果に対する表記を設定します。". Under "文体" (Style), there are three radio buttons: "自動" (Automatic), "だ・である調" (Selected), and "です・ます調" (Desu/Masu style). At the bottom, there are "キャンセル" (Cancel) and "OK" buttons.

3. テキスト翻訳とファイル翻訳



テキスト翻訳で改行を入れている例は以下のようなものがあります。改行は文区切りとみなして訳すため、文途中の不要な改行は削除した上で翻訳されることをおすすめします。

例：メール文の体裁を整えるために改行している

お世話になっております。〇〇商事のXXです。

先日ご案内いたしました新製品について、大変貴重なご意見をいただき、誠にありがとうございました。

このような文であると、改行されているため1行ずつ翻訳してしまいます

Good

先日ご案内いたしました新製品について、大変貴重なご意見をいただき、誠にありがとうございました。

改行を削除し、1文1行で翻訳すると適切に訳すことができます

例：PDFからコピー&ペーストした時、原文にレイアウト調整のための改行が含まれている

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

PDFからコピー & ペースト

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

各行の終わり等、文の途中に改行がないかを確認してください。改行があればそれを削除することで適切に訳すことができます

3. テキスト翻訳とファイル翻訳



ファイル翻訳では装飾の有無によって翻訳文が変わることがあります。もし装飾を含むファイル翻訳の翻訳精度が低いと感じた場合、一度装飾を削除して翻訳してみることをおすすめします。

• ファイル翻訳の注意点

- ファイル翻訳では、文中の装飾（太字や色、ハイパーリンク等）を保持することが可能*¹ですが、一方で、装飾の有無によって訳文が変わることがあります。

- 例：赤字にしている単語があると、対応する訳文の単語も赤字になります。
赤字にしている単語があると、対応する訳文の単語も赤字になります。

If there are words in red, the corresponding words in the translation will also be in red.
If there is a word in red, the corresponding word in the translation will also be in red.

← 翻訳文が変わっている

- 装飾が1文中に大量にあったり、意味の区切りに沿わない装飾箇所があったりすると、翻訳精度が劣化することがあります。

- 例：赤字にしている単語があると、対応する訳文の単語も赤字になります。
赤字にしている単語があると、対応する訳文の単語も赤字になります。

Word of doing in red becomes yes, and word of translation becomes red as well .
If there is a word iru in red, the corresponding translation word will also be in red.

← 精度が劣化している

ファイル翻訳の結果、精度が低いと感じたら、一度装飾を削除して翻訳してみてください。

*1 2021年9月現在で、文中の装飾保持に対応しているのは PDF, Word, PowerPoint で、Excel には未対応です。

4. ユーザ辞書とは



ユーザ辞書とは、固有名詞や専門用語をユーザが指定した通りに翻訳させることができる機能です。社内で共有されている用語集や、Mirai Translator をご使用いただく中で正しく翻訳出来なかった単語を登録していくことで、翻訳の精度を向上させていくことができます。

例

サービス開発部に所属している東海林と申します。

辞書なしでは・・・

This is Tokairin from the service development department.



人名「東海林」を誤訳している・・・
部署名「サービス開発部」はうちでは“SDD”と訳したい・・・

辞書を使うと

This is Shoji from SDD.



「東海林」が正しく訳せるようになった
うちの用語が反映された

※ ユーザ辞書を使うと、登録した単語以外の訳も変化する可能性があります。

4. ユーザ辞書とは



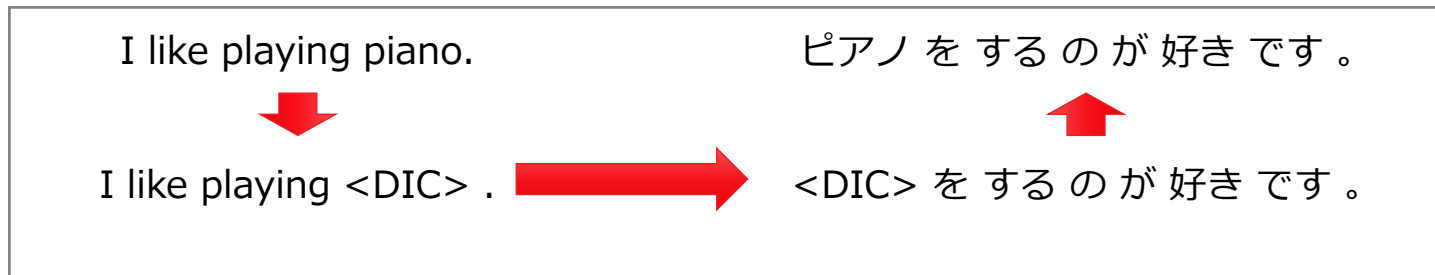
ユーザ辞書機能を適用する場合、原文の一部をマスキングした状態で翻訳することになるため、訳質の低下や誤訳につながる可能性があります。これを回避するため、1) 一般的な名詞の登録はなるべく避けること、2) ユーザ辞書なしのデフォルト設定で翻訳し、上手く訳出できない単語を順次登録することをおすすめします。

ユーザ辞書機能の仕組みと注意点

- ユーザ辞書機能は、辞書に登録されている単語を含む原文に対し、**該当部分をいったん特殊な単語（クラストークン）に置き換えてから翻訳**を行い、翻訳後に**クラストークンを辞書に登録されている訳語に書き戻す**ことで実現しています。（以下の図は「東海林 → Shoji」を辞書登録した場合）



- このため、辞書が適用されると、**原文の意味が一部欠落した状態で翻訳**をすることになり、登録する単語によっては以下の例のような誤訳が発生します。（「piano → ピアノ」を登録した場合）



* <DIC>が楽器であることが認識できていないため、play を“弾く”と訳すことが出来なくなっています（現在の機械翻訳エンジンでは、「<DIC>」は固有名詞のようなものとして認識しています）

→ 一般的な名詞はなるべく登録しないようにすることが重要です



- *1 文字正規化では、NFKC正規化を採用しており、全角の数字やアルファベットを半角に変換したりします。
- *2 翻訳前処理では、文分割処理等を行います。
- *3 トークナイズとは、文章を単語に分割することを言います。辞書のマッチングはこのトークナイズの後で行われます。
- *4 辞書に「Tokyo Tower → 東京タワー」を登録している場合