Translation Tips
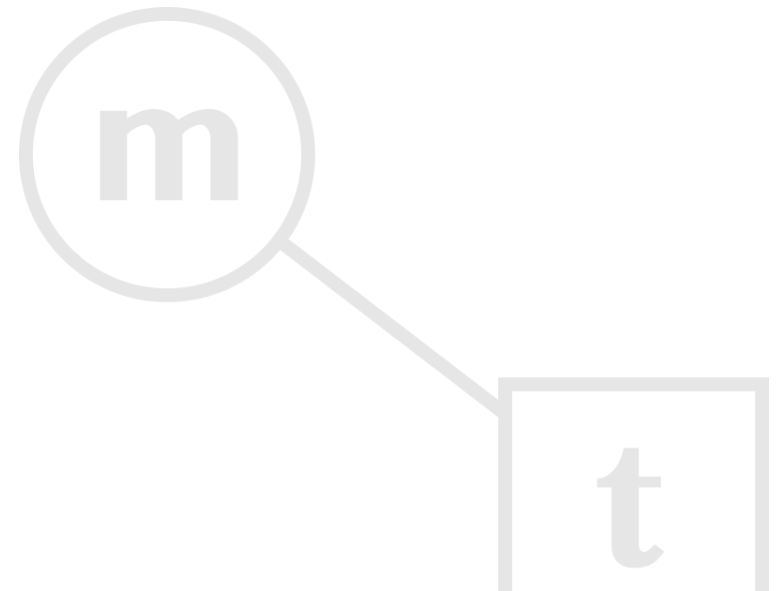
Mirai Translate, Inc.

Shibuya-Higashiguchi Bldg. 2F, 2-22-3 Shibuya,

Shibuyak-ku, Tokyo 150-0002, Japan

https://miraitranslate.com

Translation Tips

*) The content of this document is subject to change following future updates to the machine translation engine.

A type of "machine translation" called Phrase-Based Statistical Machine Translation (PBSMT) was commonly used up until the first half of the 2010s. PBSMT generates a translation by combining individual word strings, but often resulted in issues with overall fluency because sentences are composed by cutting and pasting these word strings.

Translation using SMT, the basis for the conventional technology (PBSMT)
Note: This diagram has been greatly simplified for illustrative purposes

機械翻訳　を　勉強　します　。

Phrase Table

I study　machine translation　.
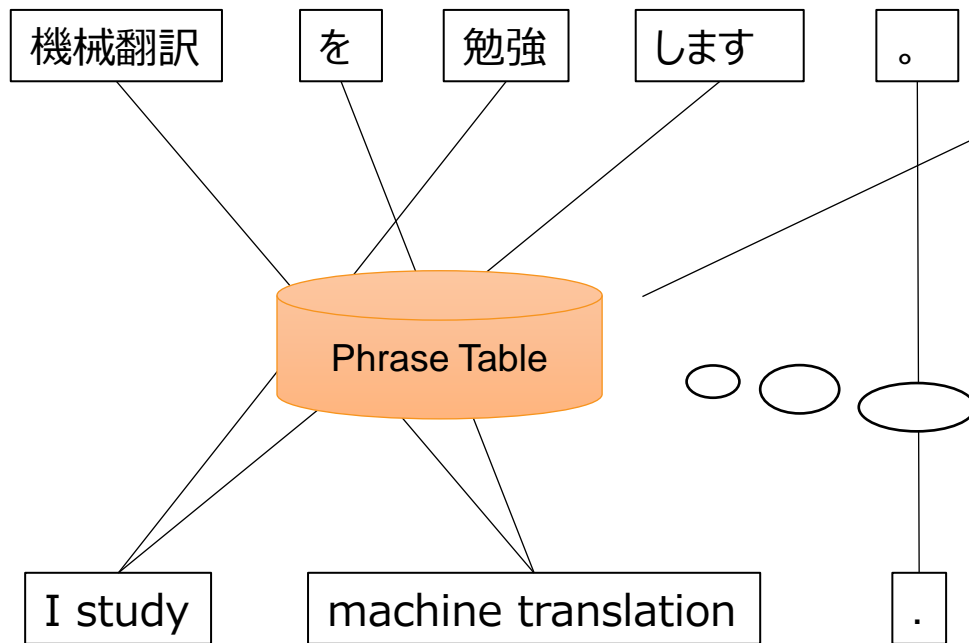
Corpus: A database built from a large range of language materials.
The bilingual corpus used in machine translation is composed of various language corpuses mapped together.

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。
:	:

The machine learns from the bilingual corpus, statistically calculates the translation with the highest probability of being used and rearranges the word order accordingly.

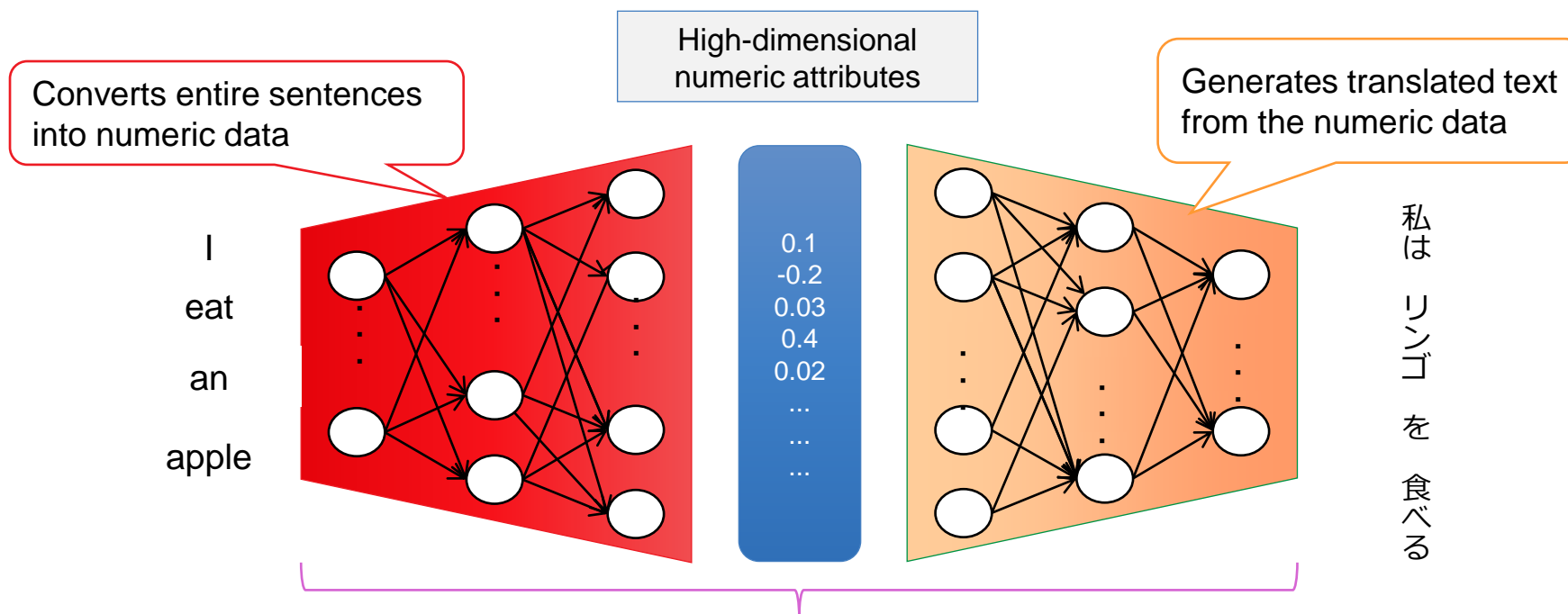**Statistical calculation concept**

	Translations for "study"	Probability of translation when "study" is used in combination with "I"
1	勉強する	80%
2	研究する	●%
3	考慮	▲%

Statistically determines the result of translating "I study" as "私は・・・勉強します".

1. Characteristics of neural machine translation

Mirai Translator uses Neural Machine Translation (NMT). Instead of working with individual word strings, NMT converts entire sentences into numeric data, and then generates a new translation from this numeric data.
As a result, one of the benefits is its ability to provide a more fluent translation result than PBSMT.

High-dimensional numeric attributes

Converts entire sentences into numeric data

Generates translated text from the numeric data

I
eat
an
apple

0.1
-0.2
0.03
0.4
0.02
...
...
...

私は リンゴ を 食べる

As the latest in artificial intelligence-type translation engines, NMT generates the translation result after converting the source text into numeric data.

Compared with conventional PBSMT technology, NMT offers dramatically improved translation accuracy.

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

Translation

PBSMT Translation result	Dinner is a Breakfast it is common in japanese inn that dinner at the dining room.

Meaning is unintelligible

Early-stage: Mirai NMT engine (2017) Translation result	There are many Japanese-style inns in the dining room for dinner.

The translation differs in meaning from the source text

Mirai NMT engine (2018) Translation result	There are many ryokans that have dinner in the room and breakfast in the dining room.

Good

Current: Mirai NMT engine (2022) Translation result	There are many ryokans where dinner is served in the room and breakfast is served in the dining room.

Very good!

The major feature provided is the high translation accuracy, and it produces overwhelmingly natural sentences compared to conventional technology.
It also has the following characteristics, and these should be kept in mind when using this technology.

Important characteristics

✓ It is difficult to tell why a given translation will go well.
-> It is not clear how to fix translation errors.

✓ The translation may not cover for any excesses or deficiencies in the source text.
-> There may be under-translation, over-translation, and/or repetition.

✓ It is sensitive to input differences, does not handle word inconsistencies well, and translation results may vary greatly depending on the presence or absence of punctuation (unreliable).

✓ It is difficult to adequately control and apply dictionaries.
-> Be sure to read "4. User dictionaries" thoroughly prior to using user dictionaries.

1. Characteristics of neural machine translation: Over-translation, Repetition, and Under-translation

NMT will sometimes generate characteristic errors that are not normally seen with conventional PBSMT technology or manual translations. When using machine translation results, it is important to check for these errors after translation.

Phenomenon	Source Text (English)	Translation Result (Japanese)
Over-translation: Outputs translated words that do not exist in the source text	The Telegraph today reported the automaker's servers have been infected with EKANS malware.	The Telegraphは米国時間8月24日、同社のサーバがEKANSマルウェアに感染したと報じた。
Repetition: Repeatedly outputs the same string	official formula	公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式公式公式公式公式公式公式公式公式 公式公式公式
Under-translation: Leaves parts of the input sentence source text untranslated	If such protective order or other remedy is not obtained, or if the Owning Party waives compliance with this Section 4.02, the disclosing party or its Affiliate or Representative, as applicable, shall (a) disclose only that portion of the Confidential Information it is compelled by law to disclose, (b) use all commercially reasonable efforts to obtain reliable assurance requested by the Owning Party that confidential treatment will be accorded such Confidential Information, **and (c) promptly provide the Owning Party with a copy of the Confidential Information so disclosed, in the same form and format so disclosed, together with a description of all Persons to whom such Confidential Information was disclosed.**	このような保護命令またはその他の救済措置が得られない場合、または所有者が本契約第4.02条の遵守を放棄する場合、開示当事者またはその関係者または代表者は、(a) 秘密情報のうち法律によって開示を強制された部分のみを開示し、(b) 所有者が要求する信頼できる保証を得るための商業的にあらゆる努力を行う。

1. Characteristics of neural machine translation: Unreliability

NMT is sensitive to input differences, does not handle inconsistencies well, and translation results may vary greatly depending on the presence or absence of punctuation.

- Presence or absence of punctuation

Example) Source text	Translation result
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。

- Notation variation

Example) Source text	Translation result
I cannot believe you.	私はあなたを信じることができません。
I can't believe you.	信じられません。

2. On writing easy to translate source text

Machine translation cannot understand the context as we humans do, so the source text must be simple and easy to understand in order to translate it correctly.
When translating English into Japanese, you can improve the accuracy of machine translation by avoiding ambiguous expressions, idioms, and spoken language, etc.

➢ Specifically, the following points apply. (English to Japanese)

1. Make sure your sentence is correct
- Make sure the spelling is correct
- Prevent unwanted line breaks

2. Use uppercase and lowercase letters properly (be careful with proper nouns in particular)
- Use uppercase letters properly
- Use lowercase letters properly

3. Use punctuation marks
- Place a period or a question mark at the end of the sentence
- Use commas for clarity

4. Eliminate Ambiguity
- Avoid anaphoric ambiguity
- Modifiers should be placed properly; next to the words they modify

5. Stay Away from Unnecessary Context-based Language
- Avoid acronyms and abbreviations
- Avoid idioms
- Avoid colloquialism(informal expression)

2. Specific examples

1. Make sure your sentence is correct

- Make sure the spelling is correct

	Example) Source text	Translation result
Wrong	It was an amature mistake.	それは大人のミスでした。
Correct (Good)	It was an amateur mistake.	それは素人のミスでした。

- Prevent unwanted line breaks

	Example) Source text	Translation result
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。
Correct (Good)	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. Specific examples

2. Use uppercase and lowercase letters properly (be careful with proper nouns in particular)

- Use uppercase letters properly

	Example) Source text	Translation result
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。

- Use lowercase letters properly

	Example) Source text	Translation result
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。

3. Use punctuation marks

- Place a period or a question mark at the end of the sentence

	Example) Source text	Translation result
Wrong	I wouldn't bet against her	私は彼女に賭けます
Correct (Good)	I wouldn't bet against her.	彼女に賭けるつもりはありません。

	Example) Source text	Translation result
Confusing	What goes up but never ever comes down	上がっても下がらない
Clear (Good)	What goes up but never ever comes down?	上がっても下がらないものは何か?

- Use commas for clarity

	Example) Source text	Translation result
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 (Good)	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。

2. Specific examples

4. Eliminate Ambiguity

- Avoid acronyms and ambiguity

	Example) Source text	Translation result
Confusing	London had snow yesterday. It fell to a depth of a meter.	ロンドン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。

> It's not a natural expression, but making the subject clear improves the translation results.

- Modifiers should be placed properly; next to the words they modify

	Example) Source text	Translation result
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かろうじて見ました。
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かろうじて見ました。

	Example) Source text	Translation result
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. Stay Away from Unnecessary Context-based Language

- Avoid acronyms and abbreviations

	Example) Source text	Translation result
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。

	Example) Source text	Translation result
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。

5. Stay Away from Unnecessary Context-based Language

- Avoid idioms

	Example) Source text	Translation result
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?

- Avoid colloquialism(informal expression)

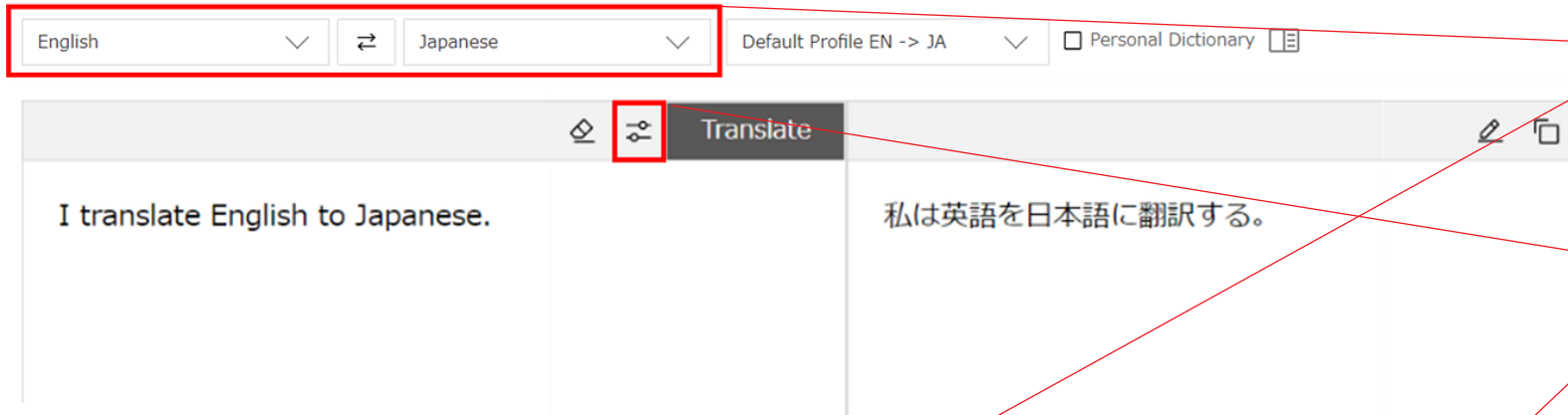|	Example) Source text	Translation result
Wrong	We saved some money on the meal so we splashed out on a really nice bottle of wine.	私たちは食事代を節約したので、本当においしいワインを一気に飲みました。
Good **Correct**	We saved some money on the meal so we spent more than usual on a really nice bottle of wine.	私たちは食事代を節約したので、本当においしいワインにいつもよりたくさんお金を使いました。

3. Text translation and file translation

There are two methods of translation: text translation and file translation. Although it is useful to use these methods differently depending on the purpose, there are certain points you should consider when using each.

- Before performing the translation
  - Make sure that the language direction is correct.
  - Translations are easier to read if you use the translation settings to unify the notation (when translating into Japanese). * By default, the notation is not unified.

- Common considerations
  - Translation is performed sentence by sentence without regard to the context (sentences cannot be translated while inferring the meaning of the surrounding sentences). Therefore, the results do not change if you enter one sentence at a time or the entire text all at once.

- Text translation considerations
  - Because line breaks are assumed to be sentence separators, inserting a line break in the middle of a sentence to align the appearance of the text (layout adjustment) will deteriorate translation accuracy.
  (Be especially careful when translating text that has been copied from a PDF, or when including multiple line breaks in email text.)

- File translation considerations
  - Decorations (bold, color, hyperlinks, etc.) may cause the translation result to differ from that of text translation.
  - When translating PDF files, some files require OCR-based character recognition. When the OCR result is poor, the translation accuracy is also more likely to deteriorate. If you have the file before converting it to PDF, applying file translation to the original file will produce more accurate results.
  - As with text translation, line breaks are assumed to be sentence separators. In particular, when translating PDF files, line breaks may occur in unintended locations during analysis.

3. Text translation and file translation

(1) Set the language direction, and in the case of translation into Japanese, (2) check whether the translation setting is appropriate.

Text Translate

| English | ⇌ | Japanese | Default Profile EN -> JA | ☐ Personal Dictionary |

Translate

I translate English to Japanese.

私は英語を日本語に翻訳する。

(1) Is the language direction for your translation correct?

(2) Which writing style will you use?

File Translate

| English | ⇌ | Japanese | Default Profile EN -> JA | ☐ Personal Dictionary |

Upload and Translate

Drag and drop files for translation

Translation Settings

Notation   File format

Set the notation for the translation result when translating to Japanese.

Writing style

○ Auto   ● Plain Form   ○ Polite Form

Cancel   OK
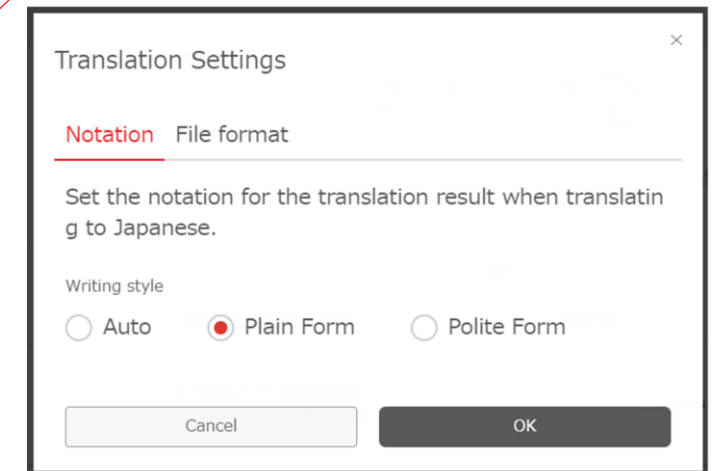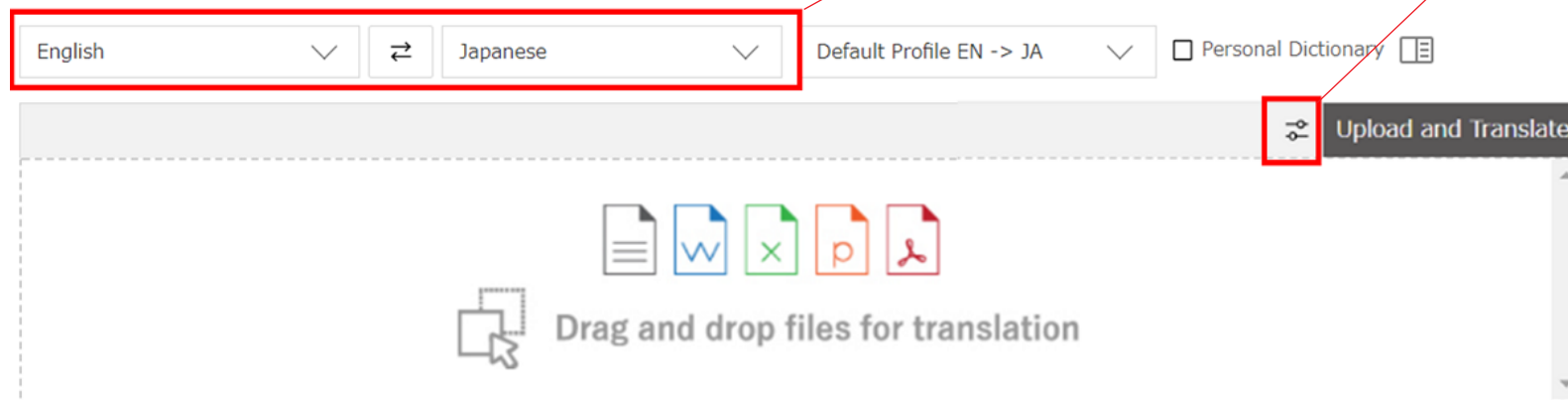
3. Text translation and file translation

The following is an example of a text translation of a source text that contains a line break. Because line breaks are assumed to be sentence separators during translation, you should delete unnecessary line breaks in the middle of sentences before translation.

e.g.: Line breaks are used to improve the appearance of the email text.

翻訳

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。

Such a statement is translated line by line because it is broken.

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。

If you delete the line breaks and translate them on a single line, you can translate them properly.

e.g.: When copying and pasting from a PDF, the source text contains line breaks for layout adjustment.

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 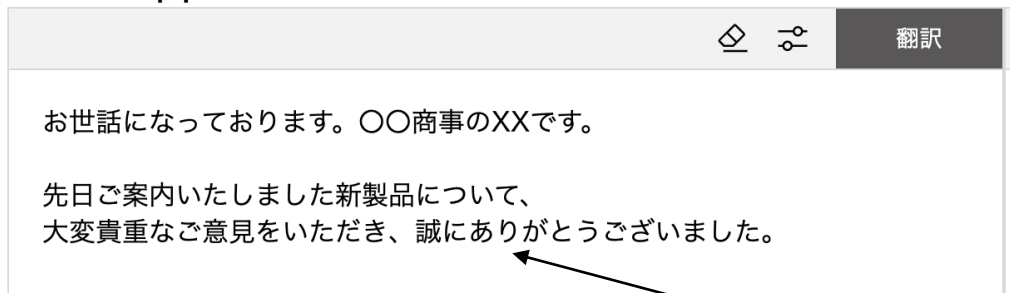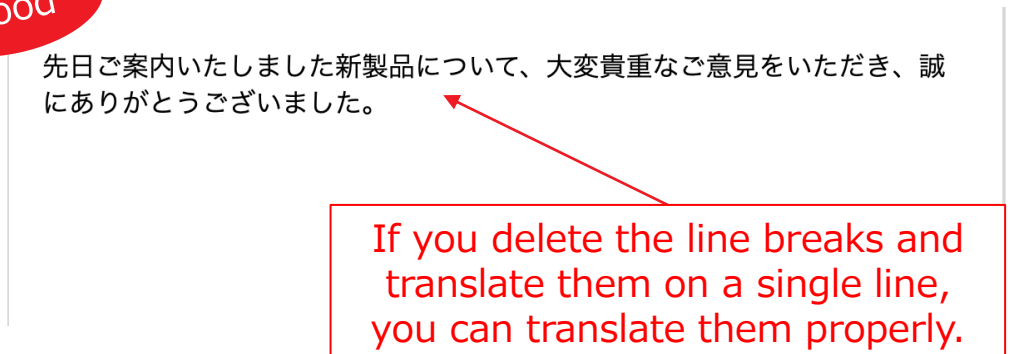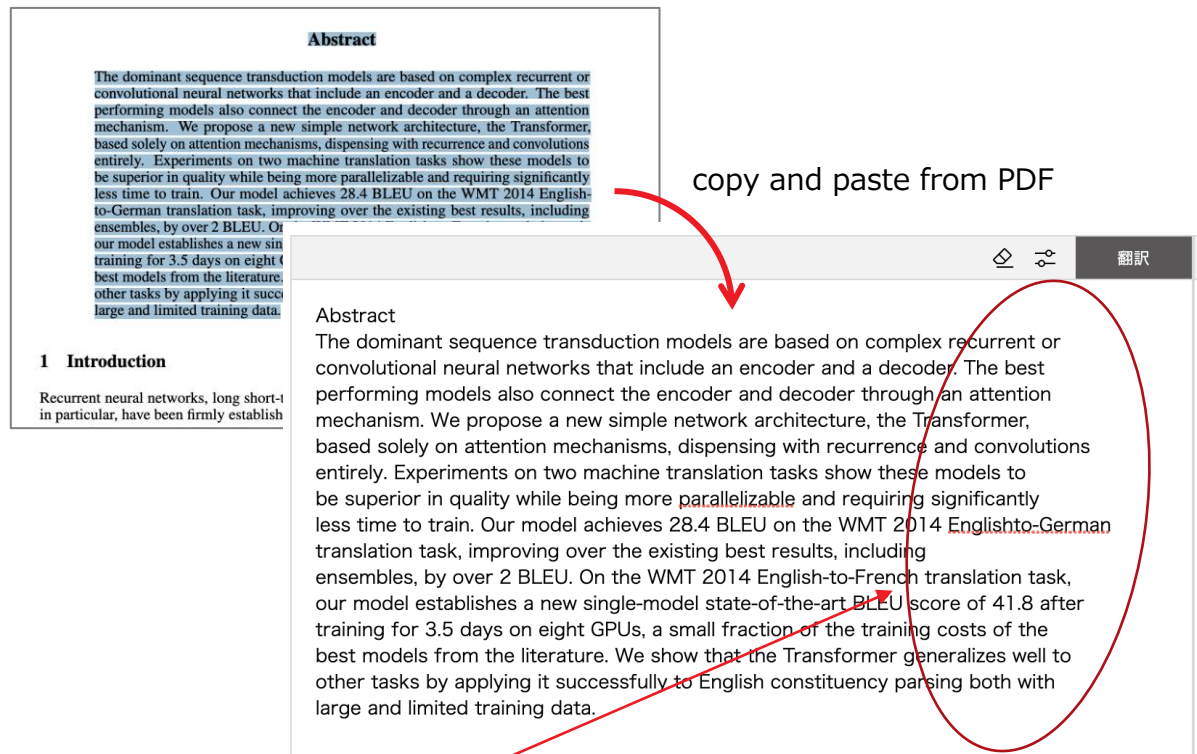solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German 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.

1 Introduction

Recurrent neural networks, long short-t in particular, have been firmly establish

copy and paste from PDF

翻訳

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 Englishto-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.

1,147/5,000

Make sure there are no line breaks in the middle of statements, such as at the end of each line.
If there is a line break, you can delete it to translate it properly.

3. Text translation and file translation

With file translation, the translation may differ depending on the presence of decorations. If you feel that the accuracy is low when translating file that includes decorations, you should delete the decorations and try translating it again.

- File translation considerations
  - With file translation, although it is possible to preserve text decorations (bold, color, hyperlinks, etc.) in sentences[*1], the translated text may change depending on the presence or absence of decorations.
    - Examples:
      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 is a word in red, the corresponding translated word will also be in red.
      If you have a word in red, the corresponding translation will also be in red. ← The translation differs

  - Translation accuracy may be deteriorated if there are a large number of decorations in a single sentence or if there are decorations that do not follow semantic separation.
    - Examples:
      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。

      The word that is in red is also in yes and the word in the corresponding translation is also in red becomes. ← The accuracy deteriorates
      If a word is in red, the corresponding translated word is also in red.

If you feel that file translation results in low accuracy, try removing text decorations prior to translation.

*1 As of September 2021, text decorations can be preserved for PDF, Word, and PowerPoint file formats. Excel is not supported.

4. User dictionaries

User dictionaries are a function that allows proper nouns and technical terms to be translated as specified by the user. You can improve the accuracy of your translations by adding glossaries that are shared within your company and words that are not translated correctly when using Mirai Translator.

Example

サービス開発部に所属している東海林と申します。

Without user dictionary

This is Tokairin from the service development department.

Mistranslated a person's name "東海林(Shoji)" as "Tokairin".
"SDD" is preferable to "service development department " as the result of translating the department name "サービス開発部 "

With user dictionary

This is Shoji from SDD.

"東海林(Shoji)" now translates correctly.
Our terminology was reflected.

*If you apply a user dictionary, translations of words besides those you have added may also change.

4. User dictionaries

When you apply the user dictionary function, a portion of the source text is translated in a masked state, which may lead to lower translation quality or mistranslations. To avoid these problems, you should 1) refrain from adding common nouns whenever possible, and 2) use the default settings without a user dictionary to translate the source text, and only add words that are not translated accurately as you come across them.

<u>User dictionary function mechanism and considerations</u>

- The user dictionary function works by first temporarily converting the corresponding part of the source text containing a word added to the user dictionary into a special word (class token) before translating the source text. After translating the source text, the system writes back the class token, using the words added as the target language, to the user dictionary. (The figure below shows what happens when "東海林 → Shoji" is added to the dictionary.)

Source text:
私は東海林です。

Convert:
私 は <DIC> です 。

Write-back:
I am Shoji.

Translate:
I am <DIC> .

- Therefore, when a dictionary is applied, translation is performed with a part of the source text meaning missing. The following example of a mistranslation can occur depending on the words you have added. (When "piano → ピアノ " is added)

I like playing piano.

I like playing <DIC> .

ピアノ を する の が 好き です 。

<DIC> を する の が 好き です 。
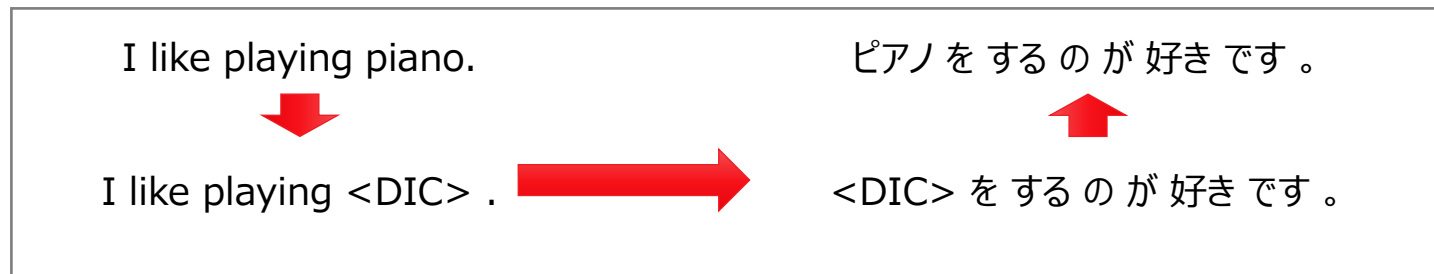
*Because the system is unable to recognize <DIC> as an instrument, it is unable to translate play as "perform".
(Today's machine translation engines recognize "<DIC>" as something similar to a proper noun)

-> It is important to avoid adding common nouns whenever possible

Reference: Flow up to translation

Input

↓

Character normalization*1

↓

Pre-translation processing*2

↓

Tokenization*3

↓

Dictionary matching*4

↓

Translation

Post-translation processing

Welcome to Tokyo Tower. GATE Ⅲ is on your right.

↓

Welcome to Tokyo Tower. GATE III is on your right.

Welcome to Tokyo Tower.

↓

Welcome to Tokyo Tower .

↓

Welcome to DICT1 .

↓

東京タワーへようこそ。

GATE III is on your right.

↓

GATE III is on your right .

↓

GATE III is on your right .

↓

ゲートIIIは右手にございます。

東京タワーへようこそ。ゲートIIIは右手にございます。

*1 Character normalization uses NFKC (Normalization Form Compatibility Composition) to convert full-width numbers and letters into half-width characters.
*2 Sentence segmentation processing is performed during pre-translation processing.
*3 Tokenization is the act of segmenting sentences into words. Dictionary matching is performed after tokenization.
*4 When " Tokyo Tower → 東京タワー " is added to the dictionary.