

Microsoft Corporation

Microsoft Translator Custom Translator User Guide

May 2018

Last Updated: May, 2018

1 CONTENTS

| | | |
|-------|---|----|
| 1 | Contents..... | 1 |
| 1. | Introduction | 2 |
| 1.1 | Audience | 2 |
| 1.2 | How This Guide Is Organized | 2 |
| 2 | Microsoft Translator Custom Translator Concepts..... | 3 |
| 2.1 | Project..... | 3 |
| 2.1.1 | The Category..... | 3 |
| 2.1.2 | The Project Label | 4 |
| 2.1.3 | Document Formats and Document Naming Convention | 6 |
| 2.1.4 | Training..... | 8 |
| 2.1.5 | Sentence Alignment in Parallel Documents | 11 |
| 3 | Building a Translation System | 11 |
| 3.1 | Custom Portal | 12 |
| 3.1.1 | Create a Project | 12 |
| 3.1.2 | Add documents..... | 14 |
| 3.1.3 | Training Custom Translator | 15 |
| 3.2 | Best Practices..... | 16 |
| 3.3 | Request Deployment | 17 |
| 4 | Appendix | 17 |
| 4.1 | FAQs..... | 17 |
| 4.1.1 | Microsoft Translator API..... | 20 |
| 4.1.2 | Glossary | 21 |

1. INTRODUCTION

Microsoft Translator Custom Translator empowers individuals, businesses and communities to build, train, and deploy customized automatic language translation systems.

Custom Translator enables the translation of specialized content to improve the quality of translations for languages supported by Microsoft Translator Neural Machine Translation. Custom also makes it possible to create translation systems for new languages that are not yet supported by the Microsoft Translator Text API service.

Custom allows you to customize a language pair for a specific domain of terminology, or to build automatic translation for a language that is not yet supported by Microsoft Translator. You can access the customized translation systems you created using the Microsoft Translator Text API, and through the applications that make use of the Translator Text API like leading translation memory applications.

Microsoft Translator is a neural machine translation system. The translation logic is stored in a *neural model*. Microsoft Translator comes with pre-built models for 100+ language pairs. Microsoft translator is used in many Microsoft products, including [Translator for Bing](#).

To create a customized translation system, you need to train a neural model. To train a model you will upload language pair documents to Custom and use them to train a model. Microsoft Translator allows you to use existing models from Microsoft in combination with your customized models, giving you wider coverage than you could achieve with your documents alone, or with Microsoft Translator alone.

You can build a translation system for a language pair that is not supported by Microsoft. You and the people you invite into your projects can create and upload training material that will be used to create a model for a new language pair.

This *User Guide* will take you through the step-by-step process of building a customized translation system using Custom.

Microsoft Translator Custom Translator is for Neural Translation. Microsoft Translator Hub is a similar system to Custom, but Hub is for Statistical Machine translation rather than Neural translation. There is no cost to use Hub. There will be a cost to use Custom. See pricing Information for Custom on the Microsoft Translator site.

1.1 AUDIENCE

This guide is for anyone interested in building a customized translation system. A deeper background in machine translation or neural networks is not necessary.

1.2 HOW THIS GUIDE IS ORGANIZED

This guide is organized into 4 major sections:

Section 1. “[Microsoft Translator Custom Concepts](#)” introduction to fundamental concepts of Custom and key terms that are used throughout this document.

Section 2. “[Building a Translation System](#)” covers topics such as how to create a project and then train your translation system. It offers guidance on the kind of language documents supported and how to select documents for tuning and testing the translation system. It introduces you to tools available in the Custom portal to evaluate the quality of the translation system. Finally, it covers the topics of deploying the customized translation system, sharing it with a broader audience and enabling people outside of your project to help improve the quality of the translation system.

Section 3. “[Appendix](#)” answers frequently asked questions and offers information on how to access the custom translation system using Microsoft Translator APIs. For novice users, it also has a glossary of commonly used terms.

2 MICROSOFT TRANSLATOR CUSTOM TRANSLATOR CONCEPTS

This section introduces you to the fundamental concepts and terminology for the Custom Translator.

2.1 PROJECT

In the Custom Portal you can create translation projects for translating from one language to another. You can create projects for the same language pair in distinct categories, and you may create projects for multiple languages. A project consists of a series of trainings, each training with its associated training documents. The language to translate from is called the *Source language*, and the language to translate to is called the *Target language*. If you are building a domain specific translation system, Custom allows you to associate a category like Sports or Medicine with your project.

2.1.1 The Category

The **category** identifies the domain type you are creating a model for. Select a category that is most appropriate and relevant to your type of documents.

You can create projects for the same language pair in distinct categories. The Custom Translator prevents creating a duplicate project with the same language pair and category, unless you also use a Label

Create New Project [X]

Project name*
Name your project(max 256 chars)

Description
Add more information to your project(max 500 chars)

Language Pair*
[Dropdown]

Category*
[Dropdown]

Category descriptor
Add a category descriptor(max 75 chars)

Project label
Add a label to your project(max 20 chars)

Create

2.1.2 The Project Label

Custom Translator allows you to assign a **Project Label** to your project. The Project Label will help distinguish one project with the same language pair and category from another project with the same language pair and category.

As a best practice, use a Project Label **only if** you are creating multiple projects with the **same** category for the **same** language pair, for instance, if you are a language service provider and want to serve multiple customers. If you use a Project Label, it is highly advisable to use the same label across language pairs, so that your application or your customer can freely switch languages and keep using the same Category ID to refer to your custom system.

The hierarchy is like this:

└ Project

The definition of a series of trainings in one language pair, one domain. The Custom Portal can have many projects. The project defines the translation system that you will use using the Translator API.

- └ Training The set of documents and the results of an individual training run for building a translation system. Your project can have many trainings. A training is a unique event in time. To reuse a training definition, you need to clone it first and then redo it. Only one training within a project is deployed and usable.

2.1.2.1 Parallel documents:

Parallel documents are pairs of documents, where one is the translation of the other. One document in the pair contains sentences in the source language and another document in the pair contains sentences in the target language, and these sentences are the same sentence in two different languages. It doesn't matter which language was the original language and which language is the translation – a parallel document can be used to train a translation system in either direction.

Parallel documents are used by the system:

- To learn how words, phrases and sentences are commonly mapped between the two languages.
- To learn how to process the appropriate context depending on the surrounding phrases. A word may not always translate to the exact same word in the other language.

As a best practice, ensure that there is a 1:1 sentence correspondence between the source and target language versions of the documents.

If your project is domain (category) specific, your documents should be consistent in terminology with that category. The quality of the resulting translation system depends on the number of sentences in your document set and the quality of the sentences. The more examples your documents contain on diverse usages for a word, the better job it can do during the translation of something it has not seen before.

You will need a minimum of 10,000 parallel sentences for full trainings. As a best practice, you can continuously add more parallel content and retrain, to improve the quality of your translation system.

Microsoft requires that documents uploaded to the Custom Translator do not violate a third party's copyright or intellectual properties. For more information, please see the [Terms of Use](#). Uploading a document using the portal does not alter the ownership of the intellectual property in the document itself.

Documents uploaded are private to each project and can be used in as many projects or trainings as you like. Sentences extracted from your documents are stored separately in your repository as plain Unicode text files and are available for you to download or delete. Do not use the Custom Translator as a document repository, you will not be able to download the documents you uploaded in the format you

uploaded them. The number of sentences extracted from each document is reported as Extracted Sentence Count in the Custom Portal.

2.1.3 Document Formats and Document Naming Convention.

File names must be at least **four** characters in length.

You can use documents in any of the following formats to build your translation system:

| Format | Extensions | Description |
|-------------------|--------------|--|
| XLIFF | .XLF, .XLIFF | A parallel document format, export of Translation Memory systems. The languages used are defined inside the file. |
| TMX | .TMX | A parallel document format, export of Translation Memory systems. The languages used are defined inside the file. |
| Locstudio | .LCL | A Microsoft format for parallel documents. The language identifier is in the file name. |
| Microsoft Word | .DOCX | Microsoft Word documents. The language identifier is in the file name. |
| Adobe Acrobat | .PDF | Adobe Acrobat portable document. The language identifier is in the file name. |
| HTML | .HTML, .HTM | HTML document. The language identifier is in the file name. |
| Text file | .TXT | UTF-16 or UTF-8 encoded text files. The language identifier is in the file name. |
| Aligned text file | .ALIGN | The extension “.ALIGN” is a special extension that you can use if you know that the sentences in the document pair are perfectly aligned. The Custom Translator will not try to align the sentences for you, it will believe your preassigned alignment. |
| Excel file | .XLSX | Excel 2013 or later document. The language identifier is in row 1 of the sheet. |

Documents can be uploaded one at a time or they can be grouped together into a zip file and uploaded.

The Custom Portal requires that you follow a **document naming convention if your file is a .zip file**:

“<document name>_<language code>.ext”

where

“document name” is the name of your document

“language code” is the ISO Language ID, indicating that the document contains sentences in that language. The language code information is displayed in the Upload Document dialog.

“ext” refers to the file name extension of the document, identifying its type.

There must be an **underscore (_)** before the language code for .zip files.

If the zip file has a nested folder structure, the Portal prefixes the folder names to the document names when it is displayed in the UI. Using folders allows you to segment your documents into distinct groups. For example, you can keep all the marketing style documents in one folder, and technical documents in another. This way you can train with or without a whole group of documents, to measure the effect of each group separately. You may also want to segment by quality. For example, if you have doubt about the origin or quality of a document, insert something in the document name that reminds you of the quality or origin. When you compose a training, you can easily filter by anything that appears in the file or folder name.

The Custom Portal does not provide an extensive file management system. As a best practice, organize and name your documents appropriately in your local file system, and then upload your documents in bulk.

Add Documents

Parallel Data

Document Type:

Training

Source (en) file:

Browse files...

Target (de) file:

Browse files...

.XLSX|.TXT|.HTML|.HTM|.PDF|.DOCX|.ALIGN file required.

Document Name*:

Name your document(max 100 chars)

or

Archive or Combo File

Combo file:

Browse files...

.TMX|.XLF|.XLIFF|.LCL|.XLSX|.ZIP file required.

Cancel

Upload

2.1.4 Training

Depending on the amount of training material you have available, one of the following three types of training is applicable to your situation:

Process: You will need to assemble training, test and tuning data for full training.

Testing and tuning data will be created for you or you can upload your own testing or tuning data.

The Training definition contains the list of documents you want to use to build your translation system, and how you want to use these documents. When setting up a training, the Custom Portal allows you to partition your documents between 3 mutually exclusive data sets. No document can appear in more than one set for this training.

1. Training data set:

Parallel documents included in this set are used by the Custom Translator as the basis for building your translation system.

You can take liberties in composing your set of training documents: Include documents that you believe are of tangential relevance and exclude them again in the next training run. If you keep the tuning set and test set constant, feel free to experiment with the composition of the training set – it is your most effective handle of modifying the quality of your translation system, after you have settled on the tuning set and test set.

As a best practice, name your documents in a way that describes their origin or quality, or arrange them in a folder on your local storage before uploading as a zip file. You can easily include them in or exclude them from a training, using the filtering option.

2. Tuning data set:

Parallel documents included in this set are used by the Custom Translator to tune the translation system for optimal results.

The tuning set is used during training to adjust all parameters and weights of the translation system to the optimal values. Choose your tuning set carefully, to be optimally representative of the content you intend to translate in the future. The tuning set has a major influence over the quality of the translations produced. Tuning enables the translation system to provide translations that are closest to the samples you provide in the tuning dataset. Only bilingual documents can be part of the tuning data set. You do not need more than 2,500 sentences as tuning set. Recommendation is to select the tuning set manually to achieve the most representative selection of sentences.

When you pick the tuning set manually, choose not too long and not too short sentences, and use sentences containing words and phrases representing the variety of words and phrases you intend to translate, in the approximate distribution that you expect in your future translations. In practice, a sentence length of 8 to 18 words will produce the best results, because these sentences contain enough context to show inflection, and provide a phrase length that is significant, without being overly complex.

A good description of the type of sentences to use in the tuning set is “prose”: actual fluent sentences. Not table cells, not poems, not lists of things, not only punctuation or numbers in a sentence - regular language.

When you let the system choose the tuning set automatically, it will use a random subset of sentences from your bilingual training documents and exclude these sentences from the training material itself. When you let the system choose the tuning set, please review it, to make sure it indeed is composed of non-trivial sentences and satisfies the criteria above.

3. Testing data set:

Parallel documents included in the testing set are used to compute the BLEU (Bilingual Evaluation Understudy) score, indicating the quality of your translation system. This score tells you how closely the translations done by the translation system resulting from this training

match the reference sentences in the test data set. The BLEU score is a measurement of the delta between the automatic translation and the reference translation. Its value ranges from 0 to 100. A score of 0 indicates that not a single word of the reference appears in the translation. A score of 100 indicates that the automatic translation exactly matches the reference: the same words are in the exact same position as the reference. The score you receive is the BLEU score average for all sentences of the testing set.

The test set should include parallel documents where the target language sentences are the most desirable translations of the corresponding source language sentences in the pair. You may want to use the same criteria you used to compose the tuning set. However, the testing set has no influence over the quality of the translation system. The Custom Translator uses it exclusively to generate the BLEU score for you, and for nothing else.

You do not need more than 2,500 sentences as the testing set. When you let the system choose the testing set automatically, it will use a random subset of sentences from your bilingual training documents and exclude these sentences from the training material itself.

You can run multiple trainings within a project and compare the resulting BLEU scores across all the training runs. You will choose to deploy the training with the best result for your production use.

During the training execution, sentences present in parallel documents are paired or aligned and the Custom Translator reports the number of sentences it was able to pair as the Aligned Sentence Count in each of the data sets. For a training run to succeed, the table below shows the minimum # of extracted sentences and aligned sentences required in each data set. Please note that the suggested minimum number of extracted sentences is much higher than the suggested minimum number of aligned sentences to consider the fact that the sentence alignment may not be able to align all extracted sentences successfully.

| Data set | Suggested minimum <i>extracted</i> sentence count | Suggested minimum <i>aligned</i> sentence count | Maximum aligned sentence count |
|-----------------|--|--|---|
| Training | 10,000 | 2,000 | No upper limit |
| Tuning | 2,000 | 500 | 2,500 |
| Testing | 2,000 | 500 | 2,500 |

The suggested minimum number of aligned sentences required is 2,000. Out of those, 1,000 aligned sentences are used in training, 500 aligned sentences are used in tuning, and another 500 are used in testing.

2.1.5 Sentence Alignment in Parallel Documents

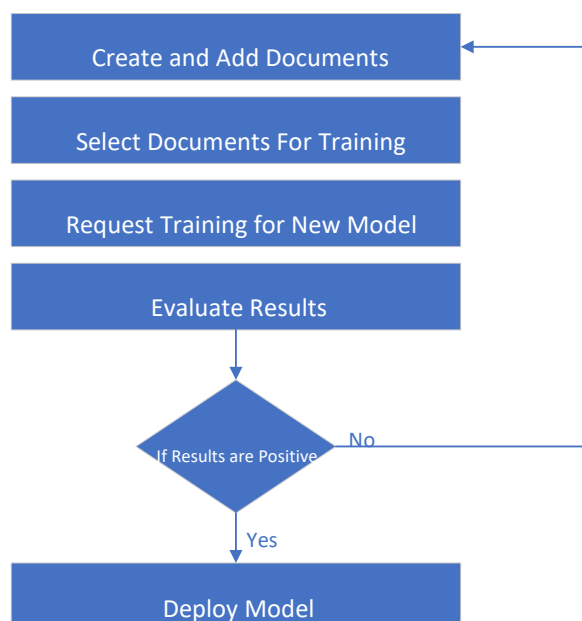
Microsoft Translator Custom learns translations one sentence at a time, by reading a sentence, the translation of this sentence, and then aligning words and phrases in these two sentences to each other. This enables it to create a map of the words and phrases in one sentence, to the equivalent words and phrases in the translation of this sentence. **Alignment** tries to ensure that the system trains on sentences that are accurate translations of each other.

The Custom Translator will automatically align the sentences in bilingual documents with the same base name you uploaded. The base name is the part of the file name before the underscore and language identifier. It will fail to do correct sentence alignment when the number of sentences in the documents differ, or when the documents you supply are in fact not 100% translations of each other. You can perform a cursory check by verifying the number of extracted sentences: if they differ by more than 5%, you may not have a parallel document.

If you know you have parallel documents, you may override the sentence alignment by supplying prealigned text files: You can extract all sentences from both documents into a text file, organize one sentence per line, and upload with an “.align” extension. The “.align” extension signals Custom Translator that it should skip sentence alignment.

3 BUILDING A TRANSLATION SYSTEM

This section takes you through a step-by-step process for **building a translation system** using the Custom Translator. The figure below is a basic flow of the process.



3.1 CUSTOM PORTAL

To use the Microsoft Translator Custom Translator Portal, you will need a Microsoft account.

When you have a Microsoft account, go to: <https://portal.customtranslator.azure.ai/> and login to the Portal.

To use the features of the Portal to train and deploy models you will need to get an Azure subscription to the Microsoft Translator Text Translation API.

Add your key to the Custom Portal by clicking on the gear icon just before your user ID in the upper right corner of the portal. Select the Add Existing Key button, enter your text translation API key and select Add. You can enter more than one key.

The custom portal encrypts all data sent to and from the Custom Translator Service. Your subscription keys are encrypted in transit and at rest.

3.1.1 Create a Project

- Select 'New Text Project'
- Give the project a name
- Enter a description
- Select a language pair
- Select a category, if you're not sure select 'General'
- Create a meaningful name for 'Category descriptor'
- Create a meaningful 'Project label'

Create New Project

×

Project name*

Name your project(max 256 chars)

Description

Add more information to your project(max 500 chars)

Language Pair*

Category*

Category descriptor

Add a category descriptor(max 75 chars)

Project label

Add a label to your project(max 20 chars)

Create

The projects page has two tabs, **Data** and **Models**. The Data tab is where you will see your documents and will be able to select documents and request the training of a model. The models tab is where you will see your models during training, when they are deployed and for Undeploying.

Projects > English to Spanish

English to Spanish [Edit](#)

Category ID: e2199527-04fd-48fa-9e39-f4a86e4dd058-3-GENERAL

Test for documentation

Language Pair: English - Spanish

Category: General

Category Description: Documentation

Project Label: 3

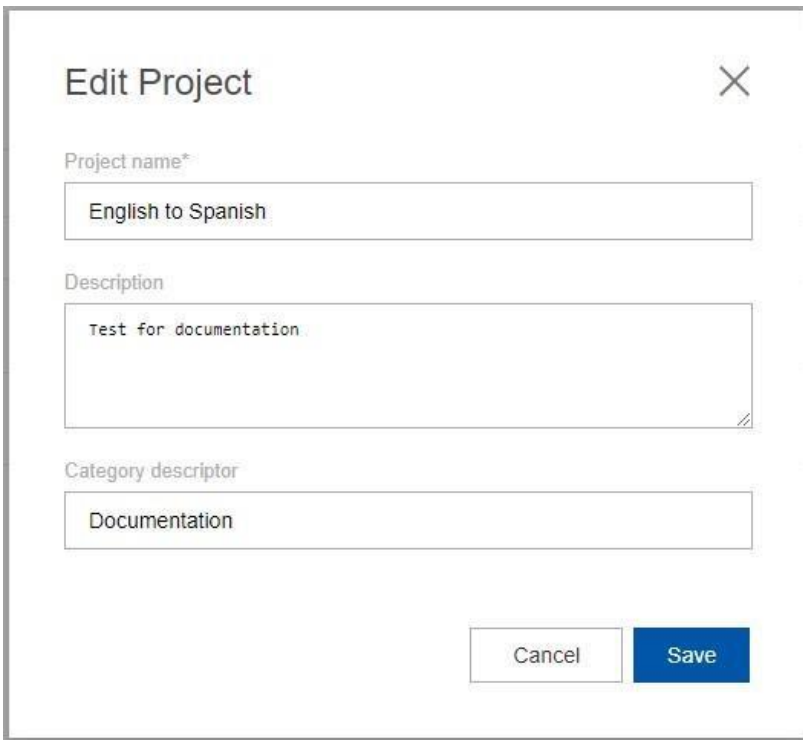
Data

Models

2 models

| Name | Status | Bleu | Training | Tuning | Test | Mono | Deploy |
|--------|----------|-------|----------|--------|-------|------|--------------------------|
| EnToEs | Deployed | 47.48 | 24,372 | 1,350 | 1,282 | 0 | Undeploy |
| Second | Running | | 24,372 | 1,350 | 1,282 | 0 | |

The projects page also contains information about your project including the category ID which is required to use a model when accessing the Microsoft Translator Text Translation API. You can edit the project by selecting **Edit** next to the name of your project.



The image shows a dialog box titled "Edit Project" with a close button (X) in the top right corner. It contains three text input fields: "Project name*" with the value "English to Spanish", "Description" with the value "Test for documentation", and "Category descriptor" with the value "Documentation". At the bottom right, there are two buttons: "Cancel" and "Save".

Project name*

English to Spanish

Description

Test for documentation

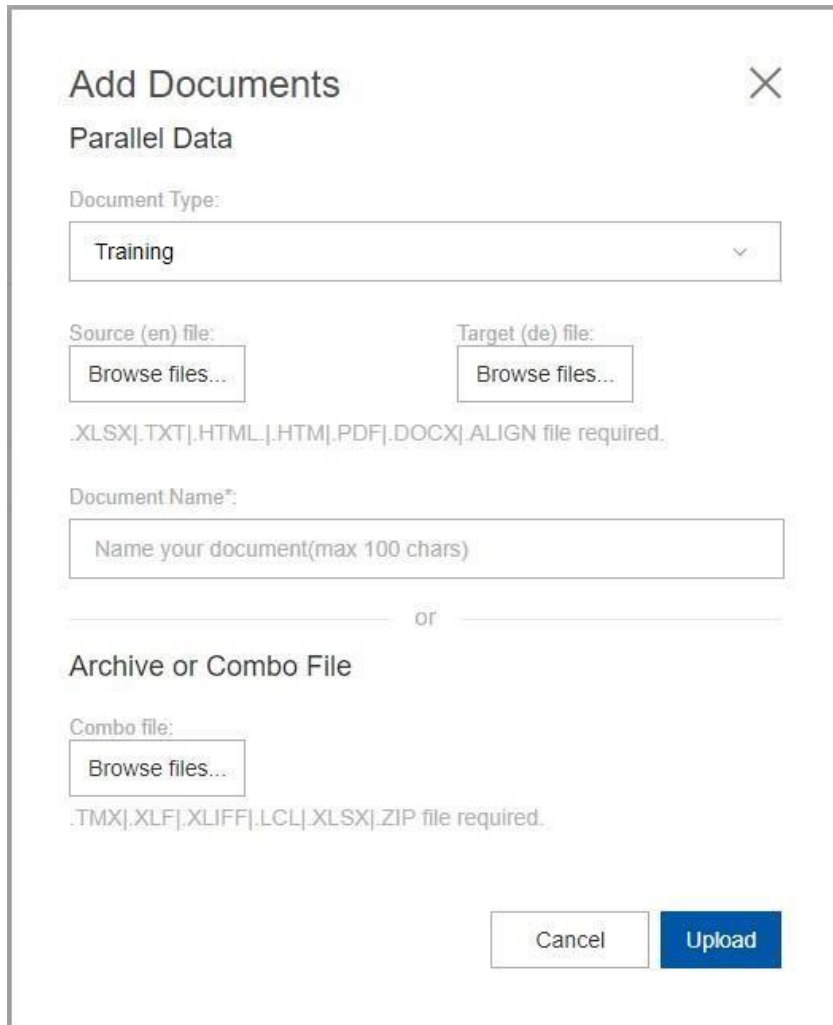
Category descriptor

Documentation

Cancel Save

3.1.2 Add documents

- Select a document type – Training, Testing or Tuning
- Select a source file
- Select a target file
- Select a meaningful document name
- OR
- Select an Archive or Combo File
- Select Upload
- If the file is .zip file then naming conventions apply. See section 2.1.4



Add Documents [Close]

Parallel Data

Document Type:

Source (en) file: Target (de) file:

.XLSX|.TXT|.HTML|.HTM|.PDF|.DOCX|.ALIGN file required.

Document Name*:

or

Archive or Combo File

Combo file:

.TMX|.XLF|.XLIFF|.LCL|.XLSX|.ZIP file required.

3.1.3 Training Custom Translator

- Select the documents you want to be used in the training.
- Select the **Train** button.
- Give the new trained model a name.
- Select the **Train model** button.
- A running model cannot be deleted. To delete a model, click on the **garbage can** icon after processing has completed.

Your new training will have a status of **Submitted** until it is accepted. The status will change to **Data processing** while the service evaluates the content of your documents.

When the evaluation of your documents is complete the status will change to **Running** and you will be able to see the number of sentences that are part of the training, including the tuning and testing sets which are created for you automatically.

You can also perform a training specifically for testing and tuning and not rely on the automatically created testing and tuning sets created for you.

Training a model can take several hours depending on the size of the files submitted and their complexity.

When a training model is complete the status will change to **Succeeded**, and a **Bleu Score** will appear. A button on the far right titled **Deploy** will appear.

ERROR CASES

3.2 BEST PRACTICES

- You can let the service automatically create testing and tuning models for the first couple of trainings for a language pair. Then download the auto-generated tuning/test set from the Training results page. Review the downloaded sentences and modify them if required. We recommend that you manually create tuning and testing datasets so that you have a steady way to compare your systems as you vary the training data to improve quality, while keeping the tuning and testing data unchanged. Both, the tuning set and the test set, should be optimally representative of the documents you are going to translate in the future.
- To compare consecutive trainings for the same systems, it is important to keep the tuning set and testing set constant. This is particularly relevant for trainings where the parallel sentence count is under 100,000 sentences. This set should be made up of sentences you think most accurately reflects the type of translations you expect the system to perform.
- If you elect to manually select your tuning data set, it should be drawn from the same pool of data as the test set, but not overlapping. There should not be a duplication of sentences between training and tuning. The tuning set has a significant impact on the quality of the translations - choose the sentences carefully.
- When you engage in a series of training runs you are likely to receive differing BLEU scores in a given language pair. Though BLEU is not a perfect metric for accuracy, there is a high likelihood that the system with a higher score provides a better translation in human judgment.
- A higher BLEU score indicates that a translation matches the target language more closely. Don't be disappointed if your scores aren't reaching 100 (the maximum). Even human translators are unable to reach 100% equivalence with each other. For ideographic and complex languages, you will get a lot of utility out of a score between 15 and 20, for most Latin script languages a score of 35 gets you into a desirable range.
- If the auto-selected sentences in the tuning data set or testing data set are not of a suitable quality in the last trained system, Custom Translator offers an option to reset the tuning data

and testing data in the new training, which forces the system to resample the tuning and testing data.

3.3 REQUEST DEPLOYMENT

It may take several trainings to create an accurate translation system for your project.

After a set of systems have been trained, go to the Project Details page and select one with a good BLEU score. You may want to consult with reviewers before deciding that the quality of translations is suitable for deployment.

If you have not already associated your Translator Text API subscription when you created your workspace, your training won't be deployed, and you will see a message to associate your Translator API subscription by clicking the **Settings** icon and adding a subscription key.

To Deploy:

- Select the **Deploy** button on the far right of the displayed model row.
- Confirm that you want to deploy.
- The **Status** of the model will change to **Deploying**.

A deployment takes much less time than training a model. The deployment should happen in several minutes.

As soon as the system is deployed, you can use it via the Microsoft Translator Text API, or any application that uses the API. Be sure to identify the correct Category ID in the API translation request.

If a model is not deployed within 90 days; then the model will be deleted.

4 APPENDIX

4.1 FAQs

Q: What are the minimum requirements for training a language pair that is not yet supported by Microsoft Translator?

A: To achieve a very basic level of understandability you will generally need 200,000 or more sentences, 10,000 sentences for each language at least Microsoft Translator CUSTOM TRANSLATOR will fail if there are less than 10,000 sentences of parallel data.

Q: When should I request deployment for a translation system that has been trained?

A: It may take several trainings to create the optimal translation system for your project. You may want to try using more training data, more or different additional target language material, or more carefully filtered data. You should be very strict and careful in designing your tuning set and your test set, to be fully representative of the terminology and style of material you want to translate. You can be more liberal in composing your training data, and experiment with different options. Request a system deployment when you are satisfied with the training results, have no more data to add to the training to improve your trained system, want to access the trained system via API's and /or want to involve your community to review and submit translations.

Q: How many trained systems can be deployed in a Project?

A: Only one trained system can be deployed per project. It may take several trainings to create a suitable translation system for your project and we encourage you to request deployment of a training which gives you the best result. You can determine the quality of the training by the BLEU score (higher is better), and by consulting with reviewers before deciding that the quality of translations is suitable for deployment.

Q: When can I expect my trainings to be deployed?

A: The deployment happens generally within 2 business days in the United States. Currently, deployments occur during the business week and there are no deployments on weekends and U.S. holidays as they require human oversight on the part of the Microsoft Translator team. On occasion there are delays if a service and/or system upgrade conflicts with deployments.

Q: How long does my system stay deployed?

A: Your system stays deployed as long as you use it. Microsoft may un-deploy the system after longer periods of non-use, typically 60 days. If you want to use the system again, simply choose the Deploy button in the CUSTOM TRANSLATOR again.

Q: How do you access a deployed system?

A: Deployed systems can be accessed via the Microsoft Translator Text API by specifying the Category ID. More information about the Translator Text API can be found at <http://www.microsofttranslator.com/dev/>

Q: How do I undeploy my translation system?

A: Please send an email to the Microsoft Translator CUSTOM TRANSLATOR Support team at customMT@microsoft.com with the name of the workspace, project and training for which the translation system needs to be undeployed.

Q: I uploaded a TMX file today for training and it exceeded the limit of 50 MB.

A: There is a 50 MB size limit for the files being uploaded. Zip the TMX file and retry the upload.

Q: The PDF file I tried to upload failed with an error saying it might be corrupt?

A: Currently, the CUSTOM TRANSLATOR cannot extract sentences from a secured PDF file. Please include only PDFs in your training that are not secured with a password.

Q: How can I ensure skipping the alignment and sentence breaking step in the CUSTOM TRANSLATOR, if my data is already sentence aligned?

A: The CUSTOM TRANSLATOR skips sentence alignment and sentence breaking for TMX files and for text files with the “.align” extension. “.align” files give users an option to the CUSTOM TRANSLATOR’s sentence breaking and alignment process for the files that are perfectly aligned, and need no further processing. We recommend using “.align” extension **only** for files that are perfectly aligned.

If the number of extracted sentences does not match the two files with the same base name, the CUSTOM TRANSLATOR will still run the sentence aligner on “.align” files.

Q: I tried uploading my TMX, but it says "document processing failed"!

A: Please ensure that the TMX conforms to the specification 1.4b.<https://www.gala-global.org/tmx-14b>

Q: How much time will it take for my training to be completed?

A: Training time depends on 2 factors: the amount of data used for training and choice of using Microsoft models. The time taken for training is directly proportional to the amount of data used to train a system. Usage of Microsoft models also increases the training time as Microsoft models are huge. Typically a training with the Microsoft models take anywhere from 4 to 12 hours to complete. Trainings without using Microsoft models may complete in less than 6 hours

Q: How does BLEU work? Is there a reference for the BLEU score? Like what is good, what the range is, etc.

A: BLEU is a measurement of the differences between an automatic translation and one or more human created reference translations of the same source sentence. The BLEU algorithm compares consecutive phrases of the automatic translation with the consecutive phrases it finds in the reference translation, and counts the number of matches, in a weighted fashion. These matches are position independent. A higher match degree indicates a higher degree of similarity with the reference translation. Intelligibility and grammatical correctness are not taken into account.

BLEU’s strength is that it correlates well with human judgment by averaging out individual sentence judgment errors over a test corpus, rather than attempting to devise the exact human judgment for every sentence.

A more extensive discussion of BLEU scores is here: <https://youtu.be/-UqDljMymMg>.

BLEU results depend strongly on the breadth of your domain, the consistency of the test data with the training and tuning data, and how much data you have available to train. If your models have been trained on a narrow domain, and your training data is very consistent with your test data, you can expect a high BLEU score. Please note that a comparison between BLEU scores is only justifiable when BLEU results are compared with the same Test set, the same language pair, and the same MT engine. A BLEU score from a different test set is bound to be different.

Q: Does the corpora need to be perfectly aligned at sentence boundaries? Though the corpora are aligned by segment, they do not always match at the sentence level. For example, a given segment might be one sentence in English, but two sentences in the target language.

A: Instances where a given segment might be one sentence in English, but two sentences in the target language, you should include them in one line and upload it as “.align” file. Sentences in “.align” file are not broken by sentence end punctuation like “.” or “;”. You can safely manage such cases via “.align” files. In “.align” files, “enter” key from keyboard is considered the end of the line/ sentence.

Q: How do I translate a local document I have on my PC?

A: Use the Microsoft Document Translator, see section 2.9.

4.1.1 Microsoft Translator API

Microsoft Translator service can be used in web or client workspaces to perform language translation and other language-translated operations. The service supports users who are not familiar with the default language of a page or workspace, or those desiring to communicate with people of a different language group.

4.1.2 Glossary

| Word or Phrase | Definition |
|--------------------------|---|
| Source Language | The source language is the language you are starting with and want to convert to another language (the “target”). |
| Target Language | The target language is the language that you want the machine translation to provide after it receives the Source language. |
| Monolingual File | This is a file containing a single language that is not paired with another file of a different language. |
| Parallel Files | This is a set of two files with corresponding text. One file contains the source language. The other contains the target language. These sentences are expected to be aligned. |
| Sentence Alignment | Parallel corpora must have aligned sentences—sentences that represent the same text in both languages. For instance, in a source parallel file the first sentence should, in theory, map to the first sentence in the target parallel file. |
| Aligned Text | One of the most important steps of file validation is to align the sentences in the parallel documents. Since things are expressed differently in different languages and different languages have different word orders, this step does the job of aligning the sentences with the same content so that they can be used for training. If your file shows a very low sentence alignment this could indicate that there is something wrong with one or both of the files. |
| Word Breaking/Unbreaking | Word breaking is the function of marking the boundaries between words. Many writing systems use a space to denote the boundary between words. Word unbreaking refers to the removal of any visible marker that may have been inserted between words in a preceding step. |
| Delimiters | Delimiters are the ways that a sentence is divided up into segments or delimit the margin between sentences. For instance, in English spaces delimit words, colons and semi-colons delimit clauses and periods delimit sentences. |
| Training Files | A training file is a file that is specifically used to “teach” the machine translation system how to map from one language (the source) to a target language (the target). The more data you can provide—either parallel or monolingual—the better the system will perform at translation. |

| | |
|--------------------|---|
| Tuning Files | These files are often randomly derived from the training set (if you select “auto”). The sentences auto-selected are used to “tune” up the system and make sure that it is functioning properly. Should you decide to create your |
| | own Tuning files, make sure they are a random set of sentences across domains if you wish to create a general purpose translation model. |
| Testing Files | These files are often “virtual” or derived files, randomly selected from the training set. The purpose of these sentences is to evaluate the translation model’s accuracy. These are sentences you want to make sure the system accurately translates. So you may wish to create a testing set and upload it to the translator to ensure that these sentences are used in the system’s evaluation (the generation of a BLEU score). |
| Translation System | When a project is created, the Microsoft Translator invokes the statistical machine translation service to build a translation system from the submitted data—either raw data that an owner has uploaded or translations submitted by reviewers. Each time a training is run within a project, a new translation system is created and access is provided to the owner to access it, review it and invite others to review its performance. |
| BLEU Score | BLEU is the industry standard method for evaluating the “precision” or accuracy of the translation model at converting text from one language to another. Though other methods of evaluation exist, this is the method that the Machine Translator Service relies on to report accuracy to Project Owners. |
| Owners | This is a person that creates the project, adds files and invites reviewers to participate in improving their translation model. Each project may have more than one owner. Owners who create the project simply invite other owners to share in the project management activities such as selecting files for the reviewers to evaluate and inviting their participation. |
| Reviewers | Reviewers are the human element of translation. They can evaluate results and test the system. Reviewers are also referred to in the workspace as “members”. |