1. Contents

1. Introduction .................................................................................................................. 3
   1.1 Audience .................................................................................................................. 3
   1.2 How This Guide Is Organized ................................................................................. 3

2. Microsoft Translator Hub Concepts ............................................................................. 4
   2.1 Workspace .............................................................................................................. 4
   2.2 Workspace Identifier ............................................................................................ 5
   2.3 Project .................................................................................................................... 6
      2.3.1 The Category .................................................................................................... 6
      2.3.2 The Project Label ............................................................................................ 6
   2.4 Parallel and Monolingual documents ..................................................................... 6
   2.5 Document Formats and Document Naming Convention ..................................... 8
   2.6 Training .................................................................................................................. 9
   2.7 Sentence Alignment in Parallel Documents .......................................................... 11
   2.8 User Roles ............................................................................................................. 12
   2.9 Collaborative Translation Framework (CTF) ......................................................... 13
   2.10 Document Translator .......................................................................................... 14

3. Building a Translation System ...................................................................................... 14
   3.1 Create a Workspace ............................................................................................... 15
      3.1.1 Associating Microsoft Translator Subscription .............................................. 16
   3.2 Setup Project .......................................................................................................... 17
      3.2.1 Create Project .................................................................................................. 17
      3.2.2 Invite Members ............................................................................................... 20
   3.3 Conduct Trainings .................................................................................................... 21
      3.3.1 Setup Training .................................................................................................. 21
      3.3.2 Using Dictionaries ............................................................................................ 32
      3.3.3 Multilingual Dictionaries .................................................................................. 33
      3.3.4 Auto-selection of Testing and Tuning data ...................................................... 33
      3.3.5 Evaluate Results .............................................................................................. 34
      3.3.6 Best Practices .................................................................................................. 36
   3.4 Share & Translate ..................................................................................................... 38
      3.4.1 Request Deployment ......................................................................................... 38
      3.4.2 Test Translations ............................................................................................. 38
   3.5 Share Translations ................................................................................................... 39
3.5.1 Engage Your Community.................................................................................................................................39
3.5.2 Use the AddTranslation() method in the Translator API .........................................................................................43
3.5.3 Re-train with Community Translations ..................................................................................................................43
4. Managing & Tracking .......................................................................................................................................................44
  4.1 Changing Workspace Name and Organization name ..................................................................................................44
  4.2 Browsing Projects .........................................................................................................................................................44
  4.3 Remove Project ..............................................................................................................................................................44
  4.4 Manage Documents .......................................................................................................................................................45
  4.5 Manage Members ..........................................................................................................................................................47
  4.6 Request Language .........................................................................................................................................................48
5. Appendix .............................................................................................................................................................................50
  5.1 FAQs ..............................................................................................................................................................................50
  5.2 Microsoft Translator API ..............................................................................................................................................56
  5.3 Glossary ..........................................................................................................................................................................57
1. Introduction

Microsoft Translator Hub empowers businesses and communities to build, train, and deploy customized automatic language translation systems – bringing better and specialized translation quality to established languages, as well as the many languages of the world that are not yet supported by major translation providers. Powered by Windows Azure™, Microsoft Translator Hub is an extension of the Microsoft Translator API and service. With Microsoft Translator Hub, businesses can build customized translations that are optimized for the company’s terminology and style, covering this company’s products and services, or the industry the organization is active in. Language communities can build translation systems that can bring better access to global information, while having the added benefit of sharing local knowledge about a culture with the rest of the world, ensuring the long-term vibrancy of languages for future generations.

*Microsoft Translator Hub allows you to customize a language pair for a specific domain (area of terminology and style) or to build automatic translation for a language that is not yet covered by Microsoft Translator. You can access the customized translation systems you created using the Microsoft Translator API, and through the applications that make use of the Microsoft Translator API, for instance most of the leading translation memory providers.*

Microsoft Translator is a statistical machine translation system, which learns how to translate from previously translated documents. The translation logic it has learnt from such previously translated documents is stored in a so-called *statistical model*. Microsoft Translator comes with pre-built models for more than 100 language pairs, which are in fact the same ones used for Bing Translator. When you train a custom translation system you build an additional statistical model just from the documents you uploaded and included in a training. Microsoft Translator allows you to use the models coming from Microsoft and your own models in combination, giving you a much wider coverage than you could achieve with your documents alone, and a much better specialization to your area of work, than the generic models from Microsoft would give you.

This *User Guide* will now take you through the step-by-step process of building your custom translation system using the Microsoft Translator Hub, referred to as the Hub from here on.

1.1 Audience

This guide will benefit any person who is interested in building a custom translation system using the Hub. A deeper background in machine translation is not essential to use the Hub.

1.2 How This Guide Is Organized

This guide is organized into following key sections

1. The first section titled “*Microsoft Translator Hub Concepts*” introduces you to the fundamental concepts of the Hub and key terms that will be used throughout this document.

2. The second section is about “*Building a Translation System*”. It covers topic such as how to create a project and then train your translation system with documents available to you. It offers guidance on the kind of documents supported and how to select documents for tuning and testing the translation system. It introduces you to the features available in Hub to evaluate the quality of the translation system. Finally, it covers the topics of deploying the translation system, sharing it with a broader audience and enabling people outside of your project to help improve the quality of the translation system.
3. The third section, “Managing & Tracking” shows how you can remove existing projects, unused documents and change roles of existing users or remove them.

4. Finally, the “Appendix” section 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 in the area of machine translation.

2. Microsoft Translator Hub Concepts

This section introduces you to the fundamental concepts of the Hub and key terms that will be used throughout this document.

2.1 Workspace

Microsoft Translator Hub provides users with a workspace to enable creation of customized translation systems.

A workspace is a work area for composing and building your custom translation system, either alone or with a community of collaborators who you can invite into your workspace. A workspace is separate from any other workspace, there is nothing that connects them. You may create or become a member of multiple workspaces, but no documents you upload are shared between workspaces, and your management of collaborators is unique per workspace. Each workspace is identified by a unique name, the Workspace Name, and has a unique ID known as the Workspace ID.

You can create a workspace simply by visiting the Microsoft Translator Hub web site. The person creating the workspace is the owner of the workspace. An Owner can then invite more people to be members of this workspace instance and can designate them either additional Owners or Reviewers. The permissions associated with each role are further discussed in Section 2.5

The concept of workspace is illustrated in Figure 1.

![Microsoft Translator Hub Diagram](image)

Documents in one workspace are not visible to users belonging to another workspace.
For example, data in Contoso’s workspace is not visible to users belonging to Hmong community’s workspace or Fabrikam’s workspace. This design isolates each workspace’s data from the other and your data is always safe.

If you are a member of multiple workspaces, you can choose the workspace to work with in a dropdown at the top right of the screen, next to your name.

2.2 Workspace Identifier

The Hub lists your workspace identifier (workspace ID) under the settings menu, between Workspace Name and Organization. The workspace ID is a fixed and constant value, assigned by the Hub when you create the workspace.

The workspace ID is the central component of the category ID, for using Hub systems programmatically via the Microsoft Translator API. See the Hub API guide for details. A Category ID consists of the workspace ID, the project label and Category code (format: Workspace ID-Project Label Category Code). The workspace ID helps you to connect Adobe Experience Manager to your custom trained translation systems.
2.3 Project

Within a workspace instance, you can create a number of translation projects for translating from one language to another. A project consists of a series of trainings, with their 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, the Hub allows you to associate a category like Sports or Medicine with your project.

2.3.1 The Category

The category identifies the domain – the general area of terminology you want to use – for your project. Please choose a category that is most appropriate and relevant to your type of documents. In some cases, your choice of the category directly influences the behavior of the Hub:

- If you choose “Technology” and choose to use Microsoft models, the Hub will use a different set of underlying models than for all other categories.
- If you choose “Speech”, the Hub will use models that are optimized for processing the output of speech recognition systems.
- Any other category selection uses the “general” Microsoft models, and is used as an identifier in your category ID.

In the same workspace, you may create projects for the same language pair in different categories. The Hub prevents creating a duplicate project with the same language pair and category. Applying a label to your project allows you to avoid this restriction. The Hub recommends to NOT use the label, unless you are building translation systems for others, for instance multiple clients.

2.3.2 The Project Label

Hub 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, in the same workspace. A project in the Hub is uniquely identified by the workspace ID, Source Language, Target Language, Category and a Project Label. As a best practice, use a 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.

2.4 Parallel and Monolingual documents

To build a translation system for your project, you need to assemble a set of documents. These documents can be of two types

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.
   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. As a result, a particular word will not always translate the same way.
As a best practice, ensure that there is a 1:1 sentence correspondence between the source and target language versions of the documents.

Example of parallel HTML documents
Description of the 2007 Office suite SP3 and of Office Language Pack 2007 SP3 in 3 languages
http://support.microsoft.com/kb/2526086/en-us
http://support.microsoft.com/kb/2526086/hi-in
http://support.microsoft.com/kb/2526086/es-es

Description of Office 2010 SP1 in 2 languages
http://support.microsoft.com/kb/2460049
http://support.microsoft.com/kb/2460049/he

Enable or disable ActiveX settings in Office files in 2 languages

Trusted documents

Example of parallel Microsoft Word documents

2. Monolingual documents:
Monolingual documents in the target language help a translation system decide which of the considered alternative translations is the most appropriate in the context, more fluent, and inflected the right way. In order for the target language document to have an effect, you also need parallel documents to generate a set of translation candidates in the first place. But even if you do not have parallel documents to generate the candidates yourself, you can expect that the underlying Microsoft models can generate the candidate, and the target language documents will help pick the right ones, if your tuning set has examples of them.

If your project is domain specific, your documents, both parallel and monolingual, 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.
We recommend that you have 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 Hub do not violate a third party’s copyright or intellectual properties. For more information, please see the Terms of Use. Uploading a document to the Hub does not alter the intellectual property in the document itself.

Documents uploaded are private to each workspace. Sentences extracted from your documents are stored separately in your workspace’s repository as plain Unicode text files, and are available for you to download. The number of sentences extracted from each document is reported as Extracted Sentence Count in the Hub.

2.5 Document Formats and Document Naming Convention.

You can use documents in any of the following formats to build your translation system in Hub.
- XLIFF (XLF, XLIFF)
- TMX (TMX)
- LCL (LocStudio generated files)
- Microsoft Word documents (DOCX)
- Adobe Acrobat files (PDF)
- HTML files (HTML, HTM)
- UTF-16 and UTF-8 encoded text files (TXT, 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 Hub skips sentence alignment on that document pair.

In order to detect parallel documents in your set, the Hub requires that you follow a document naming convention
“<document name>_<language code>.ext”
where
“document name” is the name of your document
“language code” is the standardized to ISO language IDs, indicating that the document contains sentences in that language. The language code information is displayed in the Upload Document dialog.
“ext” refers to the extension of document that should belong to one of supported document formats.

Note that there must be an underscore (_) before the language code.

Documents can be uploaded one at a time or they can be grouped together into a single zip file and uploaded. The Hub supports popular zip file formats (ZIP, GZ and TGZ).
If the zip file has a nested folder structure, Hub prefixes the folder names to the document names when it is displayed in the UI.
2.6 Training

You have following 3 options to do trainings.

Dictionary only training: You can now train a custom translation system when with just a dictionary and no other parallel documents. There is no minimum size for that dictionary, one entry is enough. Just upload the dictionary, which is an Excel file with the language identifier as column header, include it in your training set, and hit train. The training completes very quickly, then you can deploy and use your system with that dictionary. The dictionary applies the translation you provided with 100% probability, regardless of context. This type of training does not produce bleu score and this option only available if MS models are available for given language pair.

Training with 1000 parallel sentences only: You can now train a custom system with only 1000 parallel sentences. Use 500 sentences for the tuning set and 500 sentences in the test set. The Hub will build a system based on Microsoft models, and will tune the models to your tuning set, giving you a better adjusted system than the generic translation system. You may add target language documents in the desired domain as you like, the Hub will use it to build your custom target language model. It is not required though. For the Training to succeed, the 1000 parallel sentences must be unique and pass the Hub filtering. To be safe, better have 1100 or more sentences. With less than 1000 parallel sentences, you can still build a system with only a dictionary.

Full training: you will need to assemble training, test and tuning data for full training. Training is used by the Hub to understand what documents you want to use to build the translation system and how you want to use these documents. When setting up a training, Hub allows you to partition your documents between 3 mutually exclusive data sets:
1. Training data set:
Sentences of parallel and monolingual documents included in this set are used by the Hub 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. As long as 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.

2. Tuning data set:
Sentences of parallel documents included in this set are used by the Hub 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 of the documents 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 2500 sentences as tuning set. Recommendation is to select the tuning set manually, in order 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 the 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 - just 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:
The sentences of the parallel documents included in the test 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 target language 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 word is in the exact same position. 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 Hub uses it exclusively to generate the BLEU score for you, and for nothing else.

Only bilingual documents can be part of the testing data set. You do not need more than 2500 sentences as 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.

4. Dictionary (optional)
The dictionary determined the translation of phrases with 100% probability. Use it to define proper names and product names exactly the way you want to see them translated. See section Using Dictionaries for a description of the dictionary.

You can run multiple trainings within a project and compare the resulting BLEU scores across all the training runs. You would use the best training to deploy your system into production use.

During the training execution, sentences present in parallel documents are paired or aligned and Hub 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 take into account the fact that the sentence alignment may not be able to align all extracted sentences successfully.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Suggested minimum extracted sentence count</th>
<th>Suggested minimum aligned sentence count</th>
<th>Maximum aligned sentence count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>10,000</td>
<td>2,000</td>
<td>No upper limit.</td>
</tr>
<tr>
<td>Tuning</td>
<td>2,000</td>
<td>500</td>
<td>2,500</td>
</tr>
<tr>
<td>Testing</td>
<td>2,000</td>
<td>500</td>
<td>2,500</td>
</tr>
</tbody>
</table>

2.7 Sentence Alignment in Parallel Documents
Microsoft Translator Hub 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 in the other language. Alignment tries to ensure that the system trains on sentences that are actually translations of each other.

Microsoft Translator Hub 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 doing 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 pre-aligned text files: You can extract all sentences from both documents into text file, organized one sentence per line, and upload with an “.align” extension. The “.align” extension signals Microsoft Translator Hub that it should skip sentence alignment.

2.8 User Roles

There are 3 user roles in Hub

1. Owner:
   Owner is a generally the first person from the organization or community who receives an invitation from Microsoft to use the Hub and later invites other Owners and Reviewers to join. Owner has the permission to do all the activities that can be done by other Owners and by Reviewers.
   
   An Owner is authorized to do the following activities in addition to the activities that can be done by a Reviewer.
   - Create and remove projects
   - Conduct Trainings
   - Clone Trainings
   - Upload and remove documents
   - Invite Reviewers and other Owners into the workspace
   - Change the role of an existing person from Owner to Reviewer or Reviewer to Owner.
   - Assign Reviewers to review translations of sentences in the test data set.
   - Request the deployment of a translation system resulting from a successful training run.
   - Invite Community Members to post-edit sentences translated by custom translation system belonging to the workspace instance.
   - Approve/Reject alternative translations submitted for deployed translation systems.
   - Remove other Owners or reviewers from the workspace.

2. Reviewer:
   Reviewer is a person who is either invited by the Owner of the workspace.
   A Reviewer is authorized to do the following activities
   - Review translations of sentences in the test data set and submit review comments
   - Test Translations of deployed translation systems.
   - Review documents containing sentences translated by deployed custom translation system belonging to the workspace instance & submit alternate translations

3. Community Member:
   This role refers to a person who is not a part of the workspace and who has been invited by an Owner to review documents translated using the deployed translation system, and to submit alternate translations. This role is useful if you want to share selected documents containing generic sentences with a set of people without assigning them a Reviewer role.

Owners and Reviewers need to sign in with a Microsoft Account (formerly Live ID) (http://www.microsoft.com/en-us/account/default.aspx) to perform the authorized tasks. If you are using multiple Microsoft Accounts, please be sure to uncheck “Keep me signed in” in the login page, so
that you will be prompted to use the intended account with the Microsoft Translator Hub. Or invite your other accounts as members into your project.

Community Members have the option to submit alternate translations as anonymous users or authenticated users.

2.9 Collaborative Translation Framework (CTF)

The Collaborative Translation Framework is a feature of the Microsoft Translator API. It is a built-in server-side storage of sentence pairs (same sentence in source and target language). You can easily make use of the sentences stored this way, a) for immediate use in subsequent calls to the Translator API, and b) as new training data for your custom translation system. The driving idea behind the Collaborative Translation Framework is the use of the community using your application, your web site or your piece of infrastructure to provide corrections and edits of the machine translation, bringing them up to human translation quality.

If the edit was done by an authorized user having an authority of 5 or higher (In the API: rating parameter is >= 5) it will override the machine translation for this exact sentence. When you use these corrections in training, the system will learn to translate the given terminology and phrases in not exactly matching sentences as well – giving you a better targeted translation system overall.

In the Translator API, the AddTranslation() method allows you to feed a sentence pair directly.
2.10 Document Translator


Contributions to the open source project are welcome: https://github.com/MicrosoftTranslator/DocumentTranslator.

3. Building a Translation System

This section takes you through step-by-step process of building a translation system using the Hub. The figure below shows all the steps involved. Each step is further described below.
3.1 Create a Workspace

Anyone can create a workspace. To create a workspace, visit [https://hub.microsofttranslator.com](https://hub.microsofttranslator.com), and choose “Build a translation system”.

At this point, you will have to sign in with your Microsoft Account (formerly Live ID), or create a new Microsoft Account. If you have a Hotmail, Xbox or Outlook.com email address, you already have a Microsoft Account.

When logged in, you are transferred to your existing workspace, if you have one, or you get the chance to create a new one. Or you can join an existing workspace, if an owner of that workspace has invited you, by following the links in the invitation email.
3.1.1 Associating Microsoft Translator Subscription

When you create a workspace, you will have option to associate your hub workspace with your Microsoft Translator subscription. It is recommended to do this association at this stage, but if you have not subscribed to the Microsoft Translator API yet, you can do this association at a later time. Without this association, you will not be able to deploy your training or download your community translations. Use the following steps to complete the association.

1. If you don’t have subscription to Microsoft Cognitive Services follow the steps provided at [https://www.microsoft.com/en-us/translator/getstarted.aspx](https://www.microsoft.com/en-us/translator/getstarted.aspx)
2. Note down the key either of Key 1 or Key 2
3. Navigate back to [https://hub.microsofttranslator.com/Home/Settings](https://hub.microsofttranslator.com/Home/Settings)
4. Enter the subscription key that you have noted down in step 2
5. Click Save
6. Verify that the pricing tier of the subscription matches with your subscription

3.2 Setup Project

An owner can create projects. If you are building a translation system for a new language, we recommend starting with two: one for translating from a source language to the target language and one in the other direction of the same language pair. Otherwise one project is enough to start with: typically from the language your organization operates in to the target language you want to translate to. The Hub can use any bilingual material you have for building a translation system in either direction, or both directions.

3.2.1 Create Project

1. In the Projects tab, select Add Project.
2. The system shows the Create Project page.
   Enter the appropriate data into the following fields:
   - **Project Name (required):** Give your project a unique, meaningful name. It is not necessary to mention the languages within the title, though you may find that useful.
   - **Description:** Write a short summary about the project.
• **Source Language (required):** Select the language that you are translating *from.*
• **Target Language (required):** Select the language that you are translating *to.*
• **Category:** Select the category that is most appropriate for your project. The category describes the terminology and style of the documents you intend to translate.
• **Project Label (optional – not recommended to use):** 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 best practice, use a label if you are planning to build multiple trainings for same language pair, and do not use it if you are building systems for one category only. A project label is not required and not helpful to distinguish between language pairs. If you are using a label, do not add a language ID into the label.

For example, the field of Technology has a number of specific terms and requirements that require special attention. For example, a word like “cloud” may have very different meaning in the field of technology than in meteorology. If you just want to work on generic language support, choose “general”.

• **Category Descriptor:** Use this field to better describe the particular field or industry in which you are working. For example, if your category is Medicine, you might add a particular topic, such as surgery, or pediatrics. Use this to qualify the category even further. The description has no influence over the behavior of the Hub or your resulting custom system.
3. Click Create Project.

Note: You will have to type desired language to see if it is available. If a desired language is not available, click the Request a new language link. See section 4.5
Once you have successfully created a new project, you can proceed to setup a training or inviting new members.

3.2.2 Invite Members

To invite people to become members of the workspace follow the steps given below.

1. Go to the Members tab, select **Invite New Member**.
2. In the Invite Member page, enter the following information:
   - Name
   - Email Address
   - Role (select Reviewer or Owner)
   - Edit Message as needed
3. Click Send Invitation.
3.3 Conduct Trainings

3.3.1 Setup Training

1. To setup training, select the project from the **Projects** tab and go to the Project Details page. Click **Train New System**.

   **PROJECT INFO**
   
   **White Hmong to English**
   
   Started by Joe on 21 May 2022
   
   Hmong (White) to English in Medicine category
   
   You do not have any trainings to build a translation system for this language pair and category. Please click **Train New System** to setup a training.

2. This opens up the Train system page. Each system has a unique name. The Hub generates a default name which you can replace with a more descriptive name.

   **TRANSLATION SYSTEM INFO**
   
   **White Hmong to English [MSFT]**
   
   Started by Joe on 05 Oct 2013
   
   Hmong (White) to English in Medicine category
   
   **TRAIN SYSTEM** - Select/add documents and start the training
   
   **TRAIN SYSTEM**
   
   Training | Testing (Auto) | Dictionary
   
   Parallel Sentences Selected: 0 | Suggested Minimum: 10,000
   
   Parallel Documents Selected in Training Dataset: 0
   
   Monolingual Documents Selected in Training Dataset: 0
   
   Download sentence files

   Show All | Show Selected | Show Unselected

   Selected documents will be used to train your model.

   **Existing documents**
   
   **Type** | **Size (KB)** | **Sentence Count** | **Uploaded On**

   While loading this page, the Hub will search your workspace’s document repository for documents which have a language code in the document name that matches with the language code of the source or target language for this training. Multilingual documents like TMX or XLIFF files as well as the dictionary file do not need a language identifier in the name – the language ID in the XML or Excel content is sufficient.

   If documents are found, Hub will display them on this page as Parallel or Monolingual and by default include them in your training dataset. Parallel documents will be paired, so only one document name appears in the list. Please see the box titled “How Hub displays documents in the document selection grid?” for an example of how Hub displays documents in the document selection grid.

   **How Hub displays documents in the document selection grid?**

   **Assume you have the following documents**
a. If you are training an English German system, Hub will scan the workspace’s document repository for all documents with “en” and “de” in the Document name and display the following documents in the Document selection grid.
   File1.txt and File2.txt as parallel documents
   File3.txt as monolingual German document.

b. If you are training French English system, Hub will scan the workspace’s document repository for all documents with “fr” and “en” in the Document name and display the following documents in the Document selection grid.
   File1.txt as parallel document
   File2.txt as monolingual English document

c. If you are training French German system, Hub will scan the workspace’s document repository for all documents with “fr” and “de” in the Document name and display the following documents in the Document selection grid.
   File1.txt as parallel document
   File2.txt and File3.txt as monolingual German documents.

If documents are not found, you will need to upload your documents now.

Please refer to Section 2.3 for more information on Parallel and Monolingual documents and Section 2.4 for Document Naming Convention.

3. By default, the option to Use Microsoft models is checked in the Training tab, if a Microsoft model exists for this pair (approximately 100 pairs, for most languages from and to English). The effect of using this option depends on whether the source language and target language for the training are currently supported by Microsoft Translator.
   If there is no Microsoft model for your language pair, the option does not exist.

   Using Microsoft models in training the system may make your translations more accurate and more fluent. Microsoft models might not be available for some language pairs and domain combination. You can do sequential trainings with and without using Microsoft models. You will be able to get a higher score without using Microsoft models, if your training and test data are within a very narrow domain (area and of terminology and style) and will show worse results when you break out of that narrow domain. Always make sure that both your test and tuning set are representative of what you are going to translate, which is possibly less representative what you already have. In that case you will almost always get better results with Microsoft models.

4. To upload documents, click on the add documents link.

5. This opens up the Upload Documents dialog
6. Browse to the location of your documents and select the document or zip file you want to upload.

7. Click **Upload**. Hub will initiate the upload process.
8. If it is zip file, the Hub extracts all documents from the zip. For each document, it runs certain data validations and then extracts sentences from it into a sentence file.

Observe in the screenshot below, the documents named Doc005_en.pdf and Doc006_en.doc do not have a corresponding document in Hmong (mww). Hence these documents will be treated as monolingual documents for the purpose of this training, unless you upload the other document in the pair at a later point in time.

9. When the upload completes, click on Close.
10. You can also view recently uploaded documents under ‘Documents’ tab and ‘Recent Uploads’.
11. Hub will refresh the grid on the Train System page, displaying the names of the files uploaded, whether they are parallel or monolingual, the file size and sentence count for each document. For parallel documents the size and sentence count values are displayed using ‘/’ to separate the values for the source language version of the document and target language version of the document.

For example, in the screenshot below, the Hmong version of Doc0025.pdf is 58 KB in size with 58 sentences in it and the English version of Doc0025.pdf is 54 KB in size with 49 sentences in it. Each page displays 25 items. Use the paging controls to page through the list.

You will see a “New” icon next to all documents that are uploaded in the last 24 hours, to make it easier for you to add the newest documents.
12. It may happen that the sentence count for a parallel or monolingual document does not match your expectations. Hub will alert you if it detects a huge difference in the sentence count values for parallel documents.

You will see a yellow warning icon if the difference in sentence counts between the documents in the pair is greater than 10%, but less than 50% of the lower of the sentence count in the pair. You will see a red warning icon if the difference in sentence counts between the documents in the pair is greater than 50% of lower of the sentence count in the pair.

In such cases, you can select these documents in the grid and click on download sentence files link at the bottom of the grid. This will download a zip file containing the sentences extracted from each document in the pair. You can make corrections if desired and re-upload the document. You can of course also go back to the original documents, fix them, and re-upload.

If you expected a document pair to be treated as parallel documents and it isn’t, check the name of the uploaded file, make sure they have the same base name and the language identifier matching the language pair selection of this project.
13. You can also click on the sentence count value for a given document to view the sentences extracted for a given document. This is just a read-only view.

14. If the results of sentence extraction look good, you can use the checkboxes to include/exclude the documents in the training data set. If you have a lot of documents, use the search box to look for them by name. As you select the documents, Hub will update the values for “Parallel Sentences Selected”, “Parallel Documents Selected in Training Dataset” and “Monolingual Documents selected in the Training Dataset”. You will see a green tick mark next to “Parallel Sentences Selected” as soon as the no. of parallel sentences selected exceeds the suggested minimum of 10,000.

15. Integration with the collaborative features (CTF) and the Translator API: You can now directly download data you have collected from users via the AddTranslation() method in the API. Use the “get community translations” option when you are composing a training. To get these translations click the get community translations link to open up the Get Community Translations dialog. The get community translations option is available only if the training is in draft mode.
If you associated your Translator API account with your workspace, you can click ‘Get Community Translations’ to import the corrections and use them as training documents. These sentences will be stored in a file named “AlternateCommunityTranslations-<src>-<tgt>.align” as shown below.

More about the Collaborative Translation Framework in section 2.9.

16. With documents selected in the Training data set, click on the Tuning tab to select documents in the Tuning data set. By default, **Auto selection** is enabled. This allows the Hub to randomly select approximately 5% of the aligned sentences from the Training data set and consider them for tuning. For a training to succeed, Hub requires to have a minimum of 500 sentences in the Tuning data set. The Hub ensures that sentences used in the Tuning data set are not used to train the translation system.
17. If you want to manually select documents for the Tuning data set, click the **Manual selection** option. The Hub will display all documents that have not yet been selected in the Training data set and Testing data set. If you cannot find documents in this list, you should switch to either Training tab or Testing tab, unselect the desired document from that set and come back to the Tuning tab.

You should select a parallel document for which the extracted sentence count is equal. For example, in the screenshot below Doc0010.txt has 1,653 sentences in both the source language version and the target language version. Assuming the sentences in this document are well translated, sentences in Doc0010.txt are suitable for inclusion in tuning data set. In the manual selection mode, The Tuning tab currently allows you to select only one document. If you want to combine sentences from different documents, we recommend that you create a separate document containing such sentences and use it here.

To switch back to Auto selection, you can click on **auto select Tuning Data**. The Hub will automatically remove the sentences in your Tuning set from the training set, avoiding duplicates.

Even with auto selection, the tuning set remains unchanged when you add or modify the training data, until you choose to re-generate the automatically selected tuning set.

18. Having selected documents in the Tuning data set, click on the Testing tab to select documents in the Testing data set. By default, **Auto selection** is enabled. This allows Hub to randomly select approximately 5% of the aligned sentences from the Training data set and consider them for testing. For a training to succeed, Hub requires to have a minimum of 500 sentences in the Testing data set. The Hub ensures that sentences used in the Testing data set are not used to train the translation system.

19. If you want to manually select documents for the Testing data set, click the **Manual selection** option. The Hub will display all documents that have not yet been selected in the Training data set.
set and Tuning data set. If you cannot find documents in this list, you should switch to either Training tab or Tuning tab, unselect the desired document from that set and come back to the Testing tab.

In the testing data set, you should select a set of parallel documents for which the extracted sentence count is equal or almost equal. For example, in the documents selected in the screenshot below are good candidates for inclusion in the Testing data set.

To switch back to Auto selection, you can click auto select Testing Data. The Hub will automatically remove the sentences in your Tuning set from the training set, avoiding duplicates.

Even with auto selection, the tuning set remains unchanged when you add or modify the training data, until you choose to re-generate the automatically selected tuning set.

20. With the Training, Tuning and Testing data sets well defined, you are now ready to train your translation system for this language pair. To start the training, click on the Start Training button. If you want to save it and modify the training configuration, you can click Save link.

21. Before proceeding to execute the training, the Hub will alert you if it finds any issues with the documents selected in the Training, Tuning or Testing data set and also recommend fixes. You can either ignore the warning or fix the issues and then resubmit the training.
22. After your training is submitted for execution, it may take up to several hours to build your translation system depending upon the amount of data included in the training. The status will show as “In Training” on the Project Details page and you can see the more details about the Training by clicking on the System name in the Project Details page.
3.3.2 Using Dictionaries

You can specify a dictionary of terms that Microsoft Translator should use in translation, in addition to your training data for the translation system.

Use of a dictionary has the potential of degrading the quality of the translations. Here are some guidelines and hints:

- Training documents showing the terms used in context are better than a plain dictionary. Terms used in sentence form teach the system the correct inflection and agreement, better than a dictionary can.
- The dictionary maps the dictionary term or phrase exactly to the given translated form.
- Try to minimize the dictionary to the terms that the system does not already learn from the training documents.
- The dictionary works well for compound nouns like product names (“Microsoft SQL Server”), proper names (“City of Hamburg”), or features of the product (“pivot table”). It doesn’t work equally well for verbs or adjectives, because these are typically highly inflected in the source or in the target language. Avoid dictionary entries for anything but compound nouns.
- Both sides of the dictionary are case sensitive. Each casing situation requires an individual entry into the dictionary.
- You may create dictionary entries for longer phrases and expressions. In fact, entries of 2 words or longer typically have better effects on the translation than single word entries.
- Chinese and Japanese are relatively safe with a glossary. Most other languages have a richer morphology (more inflections) than English, so the quality will suffer if there is a glossary entry for a term or phrase that the system already translates correctly.

To use a dictionary, follow the steps listed here.

1. Create a dictionary of terms using Microsoft Excel.
   a. Create an Excel file. In this release, Hub supports only “.xlsx” files created using Excel 2007 and later. This file contains a list of source-language terms and a list of corresponding target-language equivalents in the first sheet of the Workbook. Other sheets in the workbook will be ignored.
   b. In cell A1 of the first sheet, enter the ISO standard language IDs for the source language (eg: “en” or “en-us” for English)
   c. In cell B1 of the first sheet, enter the ISO standard language IDs for the target language (eg: “es” or “es-es” for Spanish)
   d. Enter the source language terms in Column A, and the equivalent translations for these terms in the Target Language in Column B. HTML tags in the dictionary will be ignored.

The image below shows an Excel file containing a dictionary of terms mapped from English to Spanish.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>es</td>
</tr>
<tr>
<td>computer</td>
<td>ordenador</td>
</tr>
<tr>
<td>postcard</td>
<td>tarjeta postal</td>
</tr>
<tr>
<td>fishpond</td>
<td>vivero de peces</td>
</tr>
</tbody>
</table>

The image below shows an Excel file containing a dictionary of terms mapped from English to Spanish.
2. Save the Excel workbook.

3. Upload the dictionary file to Hub and start the training.
   a. Sign in to Hub and on the Setup Training, click the Dictionary tab.
   b. Click add documents link and upload the Excel file created in Step 1.
   c. After the upload completes, the Excel file will appear in the Dictionary tab as shown below. Select the files listed in the Dictionary tab in order to use the terms defined in those files in your training.
   d. Select Training data and click Start Training

### 3.3.3 Multilingual Dictionaries

You can create dictionaries for a specific source language and multiple target languages. To do that, simply add additional columns to the Excel sheet, as columns C, D, E and so on, following the same guidelines as above, with a language ID in row 1. Once you uploaded this Excel file into your workspace, you can use it in any project for the respective language pair, and the Hub will automatically use the appropriate language column. Column A is always the source language, and columns B to ZZ are target languages.

### 3.3.4 Auto-selection of Testing and Tuning data

Here is how auto-selection of Testing and Tuning data works in the Hub.
1. If a segment in tuning data is identical to a segment in training data, the segment is removed from training data.
2. If a segment in tuning data is identical to a segment in testing data, the segment is removed from testing data.
3. If a segment in tuning data and testing data is identical to a segment in training data, the segment is removed from training and testing data.
4. If a segment in testing data is identical to a segment in training data, the segment is removed from training data.
### 3.3.5 Evaluate Results

When the training is complete, you will receive an e-mail. You can then sign in to the Hub. Below are the steps to evaluate and act on the results.

1. **Navigate to Projects tab.**

2. **Click on a project name to open the Project Details page.** Click on the name of the system trained to open the page that shows the latest status of the training.

3. **The Status field shows the latest status for the training.** The Duration field shows how long it took to train the system starting from the time the training request was submitted till the training completed.

   If the training has succeeded, the BLEU Score field shows a value indicating how well the machine translation correlates to human judgment. For all trainings conducted after 6-July-2012, Hub shows how the trained system compares with Microsoft general domain translation system by displaying a delta of the BLEU scores between the two systems. As seen below, the score of the system trained is 8.16 points lower than Microsoft’s General Domain system for the language pair. You may consider adding more training data or train the system with Microsoft Models and see if it improves the BLEU score.

   ![Train System](TrainSystem.png)

   **Status:** Succeeded  
   **Duration:** 44 minutes  
   **BLEU Score:** 17.60 (-8.16)

4. **If the training fails, click on the Training name to view more information.** For example, you might see:

   **Status:** Training failed:
   
   Number of sentences in the training set is too small.

   This indicates that the system is not able to produce a usable translation system based on the small amount of data provided and you should find additional documents.

5. **Each of the three tabs shows the Aligned Sentence Count and Extracted Sentence Count.** Hub displays 3 sentence counts.

   a. **Extracted sentence count:** Count of sentences extracted from the document
   
   b. **Aligned sentence count:** Count of sentences aligned
   
   c. **Used sentence count:** No. of sentences used in building the system after excluding sentences that overlap with tuning or testing set.

   **NOTE:** For trainings conducted before 6-July-2012, aligned sentence count will be the same as used sentence count.

6. **If you notice a large difference between the aligned and extracted sentence counts,** you have the option to download the files and inspect them. To do so, click the checkbox next to the document(s) and click **download sentence files** to download a zip file containing the extracted sentence files and aligned sentence files corresponding to the document(s).
7. For a training that succeeds, you can click **Evaluate Results** link. This will open the Evaluate results page that shows the machine translation of sentences that were a part of the test dataset.

8. The table on the Review Translations page has two columns - one for each language in the pair. The column for the source language shows the sentence to be translated. The column for the target language contains two sentences in each row. The first sentence, tagged with "Ref:\," is the reference translation of the source sentence as given in the test dataset. The second sentence, tagged with "MT:\," is the automatic translation of the source sentence done by the translation system built after the training was conducted.

9. Click Download Translations link to download a zip file containing the machine translations of source sentences in the test data set.
This zip file contains 4 files.

a. BaselineMT_Dataset_<target language code>.txt: Contains machine translations of source language sentences in the target language done by Microsoft’s general domain system.

b. MT_Dataset_<target language code>.txt: Contains machine translations of source language sentences in the target language done by the system trained with user’s data.

c. Ref_Dataset_<target language code>.txt: Contains user provided translations of source language sentences in the target language

d. Src_Dataset_<source language code>.txt: Contains sentences in the source language

10. Navigate through the list of sentences using the paging controls and compare the machine translation against the original translation.

11. Click the Assign Reviewers to invite reviewers and other owners to review the translations for this training.

12. If you would like to retrain the system using new documents, click Clone link. This will create a copy of the existing system. You can now modify the training configuration and resubmit it.

### 3.3.6 Best Practices

- You can use the auto option for selecting a testing and tuning data set 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 switch to using the manual option for selecting tuning and testing datasets so that for over a period of time the comparisons can be done as you vary the training data 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.

- In order to compare consecutive trainings for the same systems, it is important to keep the tuning set & 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 large 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 very high likelihood that the system with a higher score would provide a better translation in human judgment. It is best to test the system or have bilingual reviewers evaluate the accuracy of a particular training.

• When you clone an existing training, we recommend that you only make changes to the training data set. In order to train a system that is comparable to the previous one, it is important to leave the tuning and testing data sets unchanged. That way, the newly trained system will be tested against the same set of sentences.

• A higher BLEU score often 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% accuracy. 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 30 gets you into a desirable range.

• When using dictionaries, try to reduce the dictionary to the terms that are absolutely necessary to regulate. In a second pass further reduce the entries to the ones that don’t come out right in a majority of test cases, and let the system choose the best option for the ones that come out right, mostly.

• If the auto-selected sentences in tuning data set or testing data set are not of a suitable quality in the last trained system, Hub 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. If you add a large number of documents to Training data set, remember to check the Reset checkbox so that the test set and tuning set are resampled. As soon as you change the test set, you have lost the ability to compare the result to a previous training. If you need to update the test set, perform a training before you make any changes, and then perform a after changing the test set and nothing else. It doesn’t matter if it lower or higher than your previous test: This is your new baseline, and you can continue improving from here.

- **Auto selection** (default)
  Sentences for tuning will be extracted from the documents selected in the Training tab. If available, sentences from the last successfully trained system will be reused. If you would like the system to use a new tuning dataset for this training, please check the "Reset" checkbox.

  [ ] Reset

• You can use the dictionary, even if you don’t use any other entries, to watermark your translation system. Put a non-word entry in it, like “mytranslatorversion”, translated to “version20150928”. You can then translate this word using your own or someone else’s application, to ensure that your application is using the correct category, and that you have reached your own custom system.
3.4 Share & Translate

This section describes how you can test the translation system you have created and then share it with people from your community so as to further improve the quality of your translation system.

3.4.1 Request Deployment

It may take several trainings in order 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.

In the Request Deployment page of the selected system, click on Request Deployment.

If you have not already associated your Translator 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 setting tab and associating your subscription.

Please allow around 2 business days for the deployment request to be processed. You will receive an email confirming the availability of your translation system as soon as the deployment has taken place. The status of a training for which a translation system has been deployed appears as “Deployed”. You can also use train and Deployment option. If you want to train and deploy system in one step. As soon as the system is deployed, you can use it via the Microsoft Translator API, or any application that uses the API. Be sure to identify the correct category ID in the API translation request.

3.4.2 Test Translations

After the translation system generated by your training has been deployed, you will receive an email from mthubsup@microsoft.com.

To verify the translation system after deployment

1. Click Projects tab.
2. From the Projects list, select the Project for which the translation system has been deployed and select the translation system with status as "Deployed". Go to the Test system page.
3. Enter the test sentence into the From box. Click Translate. The translation of the source sentence(s) will be displayed in the To box. The value to use in the category field of the Translator API, to use this deployed system, is listed on this screen.

3.5 Share Translations

After the translation system generated by your training has been deployed and you have verified it using the Test Translations page, you can share documents translated using the translation system with your community and invite them to post-edit translations to further improve the quality of the system.

3.5.1 Engage Your Community

1. Click Projects tab.
2. From the Projects list, select the Project for which the translation system has been deployed and select the translation system with status as "Deployed". Go to the Translate documents page.
3. Click Add documents and upload documents containing sentences in the source language to be translated to the target language. If the Hub finds a source language document in the repository without a corresponding target language document having the same base name, that document will be shown in the list of available documents for translation.

4. You can click on the name of the document to preview how the deployed translation system translates source language sentence in the document to the target language. The source language will be on the left, and the target language will be on the right. If you are intending to share this document with Community Members (See section 2.6), who are not a part of the workspace, please review the document to ensure that it does not contain sensitive information.
5. To share the document you have previewed, select it and click **Share**.
6. The document will now be listed under Shared Documents. Documents listed in the Shared Documents tab are visible to all Reviewers in this workspace. Reviewers can click the document name to open up the Review Documents page and submit alternate translations. To stop sharing a document, click the **stop sharing** link in the Shared Documents tab.
7. If there are a lot of documents listed in it, use the search feature to locate the document, select it and click on **Email** to invite a Community Member to review the document.
8. In the Send Email dialog, type in the email address of the community members.

```
To:* 
Subject:* Request to review documents
Message:* Hi,

Below is a list of English documents translated to Hindi which I am inviting you to review.
https://hub.microsofttranslator.com/Translation/Index/1511615?ProjectId=3079
```

9. Click **Send**.

```
Email sent successfully
```

10. Click **OK**.
3.5.2 Use the AddTranslation() method in the Translator API

To collect corrections from your users in your custom application, use the AddTranslation() method in the Microsoft Translator API. The AddTranslation() method takes a sentence pair at a time. The AddTranslationArray() method takes an array of single-sentence pairs. Whenever you use one of these methods, you are submitting this sentence pair to the CTF store.

3.5.3 Re-train with Community Translations

To re-train your translation system with community inputs, please follow the steps given below.

1. Go Projects tab
2. From the Projects list, select the Project for which the translation system has been deployed and select the translation system with status as "Deployed". Click on this training and click ‘clone’. New training will be created and you can see get community translations option under Training System tab.

3. Click Download Community Translations (under Train System tab).

4. Hub will connect to Microsoft Collaborative Translation Framework store, retrieve all the alternate translations submitted by your community and show you the Download Community Translations dialog

5. Click Download Sentence File to download a zip file containing these translations.

6. Open the zip file.
7. The downloaded zip file contains two parallel documents. Notice that these documents follow the document naming convention (Section 2.4) that is required by the Hub.
   a. ApprovedCommunityTranslations.align: This contains all sentence pairs for which translations are approved by the workspace owner/co-owner.
   b. AlternateCommunityTranslations.align: This contains all sentence pairs for which translations have not yet been approved by the workspace owner/co-owner.

8. You can review the sentences in these documents, correct them if required and re-upload them for use in your training. Since these documents already follow the naming convention, there is no need to rename the documents.

9. You can include the community submitted documents in any of the data sets – Training, Tuning or Testing. Please refer to Section 3.2.1 for further information of setting up the training.

### 4. Managing & Tracking

#### 4.1 Changing Workspace Name and Organization name

To change the name of the workspace, follow the steps given below

1. Go to **Settings** tab
2. Edit the Workspace Name or Organization Name.
3. Click the Save button.

#### 4.2 Browsing Projects

To view projects in your workspace, follow the steps given below

4. Go to **Projects** tab
5. You will see a listing of all projects in your workspace instance sorted in the reverse chronological order.
6. By default, Hub displays 5 projects per page. If you have more than 5 projects, you can select to see 10 or 15 or all projects in a single page.

#### 4.3 Remove Project

To remove a project

1. Go to **Projects** tab
2. For a selected project in the list, click on the **Remove Project** link
3. You will be presented with a confirmation dialog. Click **OK** to proceed or **Cancel** to abort.
4. Please note that removing a project cannot be undone.

### 4.4 Manage Documents

The Manage Documents page allows you to see all the documents present in your workspace’s document repository and delete documents that are not in use.

1. Go to Documents tab.
2. Hub displays Used Documents tab with a list of all documents that have either been used in a training or shared with the community for review.

3. Click **Unused Documents** tab. You will see a list of documents that have been uploaded but are not used. These documents can be safely deleted by selecting them and clicking on **Remove selected documents**.
4. The search box in at the top of the grid allows you to search by document name. You can search by the three letter language code. For e.g.: mww to see a list of all Hmong language files present in the workspace’s document repository. You can further sort by any of the column.
4.5 Manage Members

As co-owner and owner, you can manage membership of the workspace.

1. Go to **Members** tab

    1. Go to **Members** tab

    2. The system displays a list of all users of the workspace along with the roles assigned to them.

    3. As Owner/co-owner, you can select a role from the Role drop-down, change the role of an existing member and click **Save Changes**

    4. To remove an existing member, select a person from the list and click **Remove Selected Member(s)**. You will see a confirmation dialog to confirm the deletion. Click OK to proceed with it or Cancel to abort.
4.6 Request Language

MT Hub might not support the language that you want to use in your project. However, you can request a new language for your project.

1. At the time of Creating a project, if you cannot find the language in the list, select Request New Language.
2. This brings up a form

Language:* 

Tell Us More:

- About you & your organization/community:
- ISO code for the language, if one exists (eg: enu, on-US):
- Script used by the language (eg. Roman, Dravidian):
- Characters used to separate words (eg. space, comma, etc.):
- Characters used to separate sentences (eg. full stop, question mark, etc.):
- Any other names by which the language is known:

*Required

3. Enter the appropriate data into the following fields:
   - New Language (required): enter the language that is missing from the MT Hub database of translated languages.
   - Tell Us More: give as many details about the specific language, including region, dialects, etc.

4. Click on Submit Request.

Message from webpage

Your request has been submitted. You will be notified as soon as the language is available in the system.

OK
5. You should receive an email from MT Hub Support verifying your request for a new language. An email is sent to the Microsoft Translator team and they will contact your workspace owner when the language becomes available.
5. Appendix

5.1 FAQs

Q: Are the documents in my workspace visible to people outside the workspace? What is the document retention policy?
A: Information generated and stored as a result of your project from the Hub belongs to you, but may be stored for an indefinite period of time. Documents uploaded by you are only visible to authorized users in your workspace. Your data will not be shared with anyone, except the people and organizations who are tasked with the quality assurance of the Translator service. It is protected from access by anyone else. We may use your data for translation to derive statistical and quantitative information about the text being translated, in order to build and apply the most appropriate models to optimize the translation.
For more information, please see the Terms of Use.

Q: What are the minimum requirements for training a language pair that is not yet supported by Microsoft Translator?
A: Microsoft Translator Hub will fail if there are less than 10,000 sentences of parallel data. In order to achieve a very basic level of understandability you will generally need 10,000 or more sentences.

Q: What are the minimum requirements for building a category customization in a supported language pair?
A: If you are allowing the use of Microsoft models in your training setup, the minimum is the tuning set and the test set. This will give a minimal improvement for your documents at best. In order to improve further, you can add parallel material, or target language material, or both. More material is generally better, but try to stick to the category and terminology you want to improve.

Q: I selected 'Technology' category when creating a project. When I train the system for this project and check the option “Use Microsoft Models”, will the training use Microsoft’s Technology model?
A: No. At this moment the choice of “Use Microsoft models” always invokes the general training data, same as Bing Translator. The category selection serves only to identify your purpose, but has, as of now, no influence on the behavior during training or translation. That will change in an upcoming release.

Q: What if I am from a country that is not supported by the Microsoft Azure Marketplace?
A: There are a number of possibilities to work around this for your project. Please contact translator@microsoft.com for more information.

Q: What if I have documents in Word 97-2003 DOC format? Can I use them in a training?
A: You will need to use OFC.exe which is a free download included with "Microsoft Office Compatibility Pack for Word, Excel, and PowerPoint File Formats" (http://www.microsoft.com/download/en/details.aspx?id=3) and convert the DOC files to DOCX. If you have Word 2007/2010, you can use it to convert DOC files to DOCX and then upload the documents.

Q: How can I opt out of the Hub?
A: This depends on your role in the workspace you belong to. To opt out of the Hub, owners can delete themselves, while Reviewers need to request one of the owners to remove them from the workspace.
Q: When should I request deployment for a translation system that has been trained?
A: It may take several trainings in order 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 in order 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 the next business day in the United States. There will not be a deployment on U.S. holidays and weekends at this time deployment does require human oversight on the part of the Translator team. There is occasionally an additional day of delay if a system upgrade happens at the same time.

Q: After deployment, how long does my system stay deployed?
A: In order to use a customized system, you will need to deploy it, using the “Deploy” button in the Microsoft Translator Hub. Your system stays deployed as long as you use it. Microsoft may un-deploy the system after longer periods of non-use, typically between 30 and 60 days. If you want to use the system again, simply choose the Deploy button in the Hub again.

Q: Can the deployed trainings be accessed via API’s?
A: Yes. Deployed trainings can be accessed programmatically via the Microsoft Translator API (specifying the category). Details of the API can be found at the following link http://www.microsofttranslator.com/dev/

Q: How do I undeploy my translation system?
A: Please send a mail to mthubsup@microsoft.com with the name of the workspace, project & training for which the translation system needs to be undeployed.

Q: After deployment of the trained system, there does not seem to be a way to upload a TMX file and get it machine translated on the server side?
A: Most of the commercial and open source TM tools offer a way to translate TMX files using a custom MT system. Microsoft does not directly offer a TMX translation tool.

Q: I trained and deployed a customized MT system last week, made a few translations, and this week I notice that my translations are different than last week.
A: Microsoft regularly updates the base translation models that are being used in all translations using Microsoft models, and the implementation of the translation engine itself. It may happen every couple
of weeks. Each one of these basic changes has the chance to change the translations being produced with your custom system. This is not a cause of concern, because Microsoft makes sure that the majority of changes are positive. If you notice your translations have changed, and you want to verify if they are positive even for your own test set, you may go ahead and request a new training, which will produce a new BLEU score, based on the new Microsoft models. In more cases than not, the new score will be slightly higher than the previous one. It may happen that in particular cases the score is lower. The only certain thing is that on average the scores will be higher. The changed translations are unavoidable when upgrading the basic translation models and algorithms.

Q: TMX file fails to upload with an unknown language error message?
A: This happens when the TMX files actually has wrong language code as “ES-EM” instead of “ES-ES” or does not have the right format “en_US” opposed to “en-US” as expected by MT Hub.

Q: I uploaded a TMX file today for training and I got the message that the file size exceeded the limit of 50 MB.
A: Yes we do have a 50 MB size limit for the files being uploaded. Zip the TMX file and retry the upload.

Q: My training failed! How can I avoid my trainings from failing?
A: Trainings can fail if they do not meet the constraints for the minimum required sentences in the Test, Training or Tuning data. The number of minimum required aligned sentences for a Training to succeed is 500 for the Tuning and Testing set. For the training set, it is 2000.

If your training fails with the message “An error occurred while building the translation system. Please try again after some time. If you continue to get the same error, please email mthubsup@microsoft.com.” we recommend to wait for few hours before re-submitting the system for training. If you are encountering these errors on a regular basis and the Hub team has not already reached out to you, please send an email to mthubsup@microsoft.com.

Q: How can I download community translations?
A: Please refer to section 3.3.3.3 of the user guide.

Q: Is there a feature in MT HUB which would enable a project owner to approve all the submitted translations?
A: Translations provided by the community or reviewers can be approved all at once. To approve the translations, navigate to “Community > Invite Reviews”, click on the “Manage Translations” link and select the “Suggested” radio button on the Manage Translations page. Please refer to section 3.3.3.2 of the user guide

Q: The PDF file I tried to upload, failed with an error saying it might be corrupt?
A: The PDF file that failed to upload may be a secure PDF file. Currently Hub 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 MT Hub, if my data is already sentence aligned?
A: MT Hub skips sentence alignment and sentence breaking for .tmx files and for text files with “.align” extension. “.align” files give users an option to skip MT Hub’s sentence breaking and alignment process
for the files that are perfectly aligned, and need no further processing. We do 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 Hub will still run the sentence aligner on “.align” files.

Q: Is there a way to upload a TMX file and get it machine translated on the server side?
A: The machine translations can be viewed via the test console or can be retrieved via an API. We do not currently offer a direct TMX translation utility. Many commercial TM tools offer TMX translation.

Q: Is there an option to have multiple instances of the same language/category combination?
A: As a workaround for now we can provide you with another instance of a workspace, so you can create a project for the required language/category combination and the data will also remain isolated. To request another instance, reach out to MTHubSup@microsoft.com.

Q: I tried uploading my TMX, but it said "document processing failed"!
A: Please ensure that the TMX conforms to the specification 1.1 or 1.4b
http://www.localization.org/tmx/tmx.htm

Q: How much time will it take for my training to complete?
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 Microsoft model would take anywhere from 4 to 12 hrs to complete. Trainings without Microsoft model may complete in less than 6 hrs.

Q: Why the results from the “Test Translation” page of Microsoft Translator Hub differ from the one returned by the Microsoft Translator API with Hub? Is it the difference from the two content types of "text/plain" and "text/html"?
A: Yes the web interface in the Hub uses contentType="text/plain". In plain text, tags that look like <one letter><number> are left untouched and move with the word they are next to. This may result in tag ordering that would be illegal in XML. Tags of other format will not be treated as tags. The Hub forces all tags it sees in the sample documents into the <one letter><number> format, but the API won’t.
In text/html proper HTML processing is done, tags will be in legal order and legal nesting. However, you must pass balanced HTML, and self-closing tags will be expanded in the process. You will want to use text/plain for most content, except when you have balanced HTML, or balanced XML that you can transform HTML. In contentType=text/html you may also exclude any span of text from translation by using the nottranslate attribute.
When using HTML, the engine does a better job at positioning the tags properly. If you use plain text and have tags in there, you will need to ensure the correct tag placement yourself.

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](https://youtu.be/-UqDljMymMg).

All that being said, 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: Do 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 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 “;”. Hence 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: Uploading a gz file gives an error: “The document has no extension. Please upload a document with a supported file extension. “
A: Certain version of gz files is not supported by MT Hub gz extractor. The workaround is to create a new gz file in 7Zip.

Q: How do I translate a local document I have on my PC?
A: Use the Microsoft Document Translator, section 2.10.
<table>
<thead>
<tr>
<th><strong>GZIP archive</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version to extract</strong></td>
<td>Unknown</td>
</tr>
<tr>
<td><strong>Host OS</strong></td>
<td>UNIX</td>
</tr>
<tr>
<td><strong>Total files</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total length</strong></td>
<td>100,506,704</td>
</tr>
<tr>
<td><strong>Packed length</strong></td>
<td>21,301,849</td>
</tr>
<tr>
<td><strong>Ratio</strong></td>
<td>21%</td>
</tr>
<tr>
<td><strong>SPX module size</strong></td>
<td>0</td>
</tr>
<tr>
<td><strong>Main comment</strong></td>
<td>Absent</td>
</tr>
<tr>
<td><strong>Passwords</strong></td>
<td>Absent</td>
</tr>
</tbody>
</table>
5.2 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.

Documentation for the Translator APIs can be found at the [MSDN Translator API](https://docs.microsoft.com/en-us/translator/api) article. The same API is available for customized translation system. Please see the [MT Hub API Guide](https://docs.microsoft.com/en-us/translator/api) for further details on how to use Microsoft Translator APIs to access your translation system.
### 5.3 Glossary

<table>
<thead>
<tr>
<th>Word or Phrase</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual File</td>
<td>This is a file containing a single language that is not paired with another file of a different language.</td>
</tr>
<tr>
<td>Parallel Files</td>
<td>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.</td>
</tr>
<tr>
<td>Source Language</td>
<td>The source language is the language you are starting with and want to convert to another language (the “target”).</td>
</tr>
<tr>
<td>Target Language</td>
<td>The target language is the language that you want the machine translation to provide after it receives the Source language.</td>
</tr>
<tr>
<td>Sentence Alignment</td>
<td>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 there should, in theory, map to the first sentence in the target parallel file.</td>
</tr>
<tr>
<td>Training Files</td>
<td>A training file is a file that is specifically used to “teach” the machine translation system on 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.</td>
</tr>
<tr>
<td>Tuning Files</td>
<td>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.</td>
</tr>
<tr>
<td>Testing Files</td>
<td>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).</td>
</tr>
<tr>
<td>Translation System</td>
<td>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.</td>
</tr>
<tr>
<td>BLEU Score</td>
<td>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 method that the Machine Translator Service relies on to report accuracy to Project Owners.</td>
</tr>
<tr>
<td>Owners</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Reviewers</strong></td>
<td>Reviewers are the human element of translation. They can view the translations, rate them and provide suggested alternatives or corrections. Reviewers are also referred to in the workspace as “members”.</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Aligned Text</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Word Breaking/Unbreaking</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Delimiters</strong></td>
<td>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.</td>
</tr>
</tbody>
</table>