AutoML Translation trains custom models using matching pairs of sentences in the source and target languages. It treats each sentence pair as an independent training item, without assuming any correlation between separate pairs.

The sentence pairs used to train your custom model must be in Tab-separated values (.tsv) or Translation Memory eXchange (.tmx) (https://en.wikipedia.org/wiki/Translation_Memory_eXchange) format. You can batch multiple .tsv and .tmx files into a comma-separated values (.csv) file.

You can import individual .tsv or .tmx files using the AutoML Translation UI; the AutoML API supports only .csv files.

AutoML Translation uses the sentence pairs you provide to train, validate, and test the custom model.

- `TRAIN` - Use the sentence pairs to train the model.
- `VALIDATION` - Use the sentence pairs to validate the results that the model returns during training.
- `TEST` - Use the sentence pairs to verify the model's results after the model has been trained.

Optionally, you can control which sentence pairs AutoML Translation uses for each purpose by uploading separate training, validation, and testing files. If you don't explicitly specify which files to use for these three purposes, AutoML Translation automatically divides your sentence pairs into three sets. AutoML Translation uses approximately 80% of your data for training, 10% for validation, and 10% for testing (up to 10,000 pairs for validation and testing).

don't explicitly specify training, validation, and test sets, the minimum number of sentence pairs to tra
l is 1000. If you exceed 100,000 sentence pairs, AutoML Translation only picks 10,000 sentence pairs
ation and test sets each, respectively. Other sentence pairs are pushed to training set. In this case, you
an unbalanced data split which can have an impact on your model. AutoML Translation provides a
ng message but still allows training to proceed. If you choose to manually specify training, validation,
ets, then validation and test sets can't exceed 10,000; otherwise, AutoML Translation rejects the impor
s an error.

Sentence pairs are always de-duplicated across all imported sentence pairs. A sentence pair is a duplicate of another when their source sentence matches another source sentence. In addition, AutoML Translation doesn't allow you to import 2 file with same content being imported twice.

You must provide at least three sentence pairs for data labeled `TRAIN`. The maximum supported size of a dataset is 15M. The minimum number of sentence pairs used for VALIDATION or TEST is 100 each. The maximum number of pairs used for VALIDATION or TEST is 10,000;

Among multiple data imports of the same dataset, you can specify training/validation/test sets for one import and use automatic split for another.

Data is always re-balanced with respect to your manual division after each import and file deletion.

AutoML Translation supports tab-separated files, where each row has this format:

- `Source sentence` *tab* `Translated sentence`

For example:

All text in a .tsv file must be plain text. If the text includes HTML tags or other markup, AutoML Translation treats the markup as plain text.

The tab-separated source data does not include language codes to identify the source and target languages. You identify the source and target language codes when you describe the model to be trained. AutoML Translation interprets the first segment as the source language, the second segment as the target. In the example above, the source would be English, and the target would be German.

Translation Memory eXchange (TMX) is a standard XML format for providing source and target translation sentences. AutoML Translation supports input files in a format based on TMX version version 1.4. This example illustrates the required structure:

The <header> element of a well-formed `.tmx` file must identify the source language using the `srclang` attribute, and every <tuv> element must identify the language of the contained text using the `xml:lang` attribute.

All <tu> elements must contain a pair of <tuv> elements with the same source and target languages. If a <tu> element contains more than two <tuv> elements, AutoML Translation processes only the first <tuv> matching the source language and the first matching the target language and ignores the rest. If a <tu> element does not have a matching pair of <tuv> elements, AutoML Translation skips over the invalid <tu> element.

AutoML Translation strips the markup tags from around a <seg> element before processing it. If a <tuv> element contains more than one <seg> element, AutoML Translation concatenates their text into a single element with a space between them.

If the file contains XML tags other than those shown above, AutoML Translation ignores them.

If the file does not conform to proper XML and TMX format – for example, if it is missing an end tag or a <tmx> element – AutoML Translation aborts processing it. AutoML Translation also aborts processing if it skips more than 1024 invalid <tu> elements.

To upload sentence pairs using the AutoML API, you create a comma-separated values (.csv) file that identifies the .tsv and .tmx files to use, and which can also indicate which pairs to use for training, validation, and testing. The .csv file can have any filename, must be UTF-8 encoded, and must end with a .csv extension. The file has one row for each .tsv or .tmx file you are uploading, with two columns in each row:

- Which set to assign the sentence pairs in this file to. This field is optional and can be one of these values:
  - TRAIN
  - VALIDATION
  - TEST
  - UNASSIGNED

    If a dataset is specified as UNASSIGNED, then AutoML Translation automatically splits it to ensure that there is enough training, validation, and testing content.
- The full path to a .tsv or .tmx document containing sentence pairs.

For example, you might have the following in your .csv file: