Document Translation

Overview

Document Translation is a new feature in [Azure Translator service](#) which enables enterprises, translation agencies, and consumers who require volumes of complex documents to be translated into one or more languages preserving structure and format in the original document. It asynchronously translates whole documents in a variety of file formats including **Text, HTML, Word, Excel, PowerPoint, Outlook, PDF, and Markdown** across any of the <mark>111 languages and dialects</mark> supported by Translator service.

Standard translation offerings in the market accept only plain text or html, and limits count of characters in a request. Users translating large documents must parse the documents to extract text, split them into smaller sections and translate them separately. If sentences are split in an unnatural breakpoint it can lose the context resulting in suboptimal translations. Upon receipt of the translation results, the customer must merge the translated pieces into the translated document. This involves keeping track of which translated piece corresponds to the equivalent section in the original document. The problem gets complicated when customers want to translate complex documents having rich content.

Document Translation makes it easy for the customer to translate:

a) volumes of large documents,
b) documents in variety of file formats,
c) documents requiring preserving the original layout and format, and
d) documents into multiple target languages.

User experience

User makes a request to the Document Translation service specifying location of source and target documents, and the list of target languages. The service returns an identifier enabling the user to track the status of the translation. Asynchronously, Document Translation pulls each document from the source location, recognizes the document format, applies right parsing technique to extract textual content in the document, translates the textual content into target languages. It then reconstructs the translated document preserving layout and format as present in the source documents, and stores translated document in a specified location. Document Translation updates the status of translation either at the job or document level.

Users can provide a custom model id built using custom translator portal, custom glossaries, or both as part of the request to translate documents. Document translation applies such customization retaining specific terminologies and providing domain specific translations in the translated documents.