

Evaluation Report for Customization of Amazon Translate Active Custom Translation with TAUS Medical/Pharma Data

Language Pair

English → Swedish

Domain

Medical/Pharma

Introduction

Online machine translation engines provide easy access to high quality machine translations. These machine translation engines are optimized for content like news articles and social media posts that end users of these online platforms frequently translate.

Businesses often want to translate text with a different style and a specific topic. For enterprise use, online machine translation engines offer customization with existing translations that reflect the desired style and topic.

TAUS makes such customization data available via the TAUS Data Marketplace and TAUS Matching Data platforms, and now AWS Marketplace.

So that business can clearly assess the value of TAUS data, Polyglot Technology LLC was tasked to independently evaluate the quality of machine translation of Amazon Translate customized with TAUS Data (using [Amazon Translate Active Custom Translation](#)) compared to non-customized Amazon Translate.

BLEU Scores for Amazon Translate and Amazon Translate Active Custom Translation

Machine Translation Evaluation

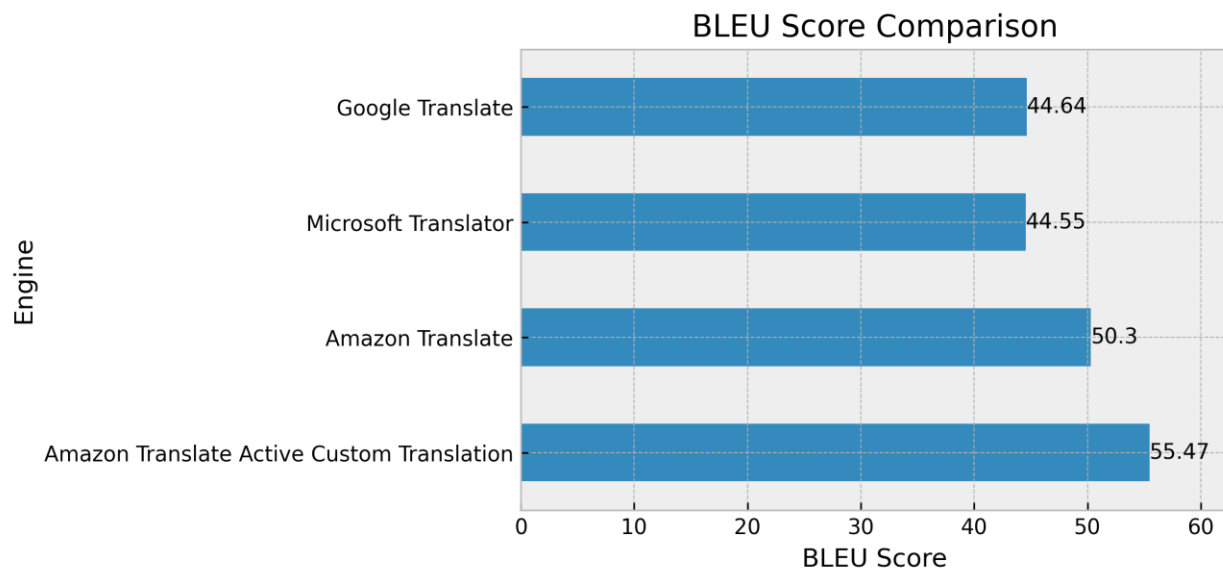
To judge whether machine translation is good or not, human evaluation is the best method. We can ask speakers of the source language and the target language, or better professional translators, to judge whether a machine translation is an adequate and fluent translation of the original text. Or we can ask how close the machine translation is to a human reference translation. Human evaluation however, is slow and hard to scale along language pairs and domains.

Automatic metrics that also use human reference translations have been developed to calculate a numeric score for machine translation quality. For close to 20 years the predominant automatic metric is BLEU, measuring the similarity of machine translations to human reference translations on a scale from 0 to 1 (or 0 to 100 when expressed as percentages). More details on BLEU and how to interpret it can be found in the section “Interpreting BLEU Scores” below.

TAUS Test Set

TAUS selects the machine translation customization data by querying its large repository of high-quality translation data with a domain-specific text. The resulting customization dataset is then split at random into a larger training set for Amazon Translate Active Custom Translation and a smaller 2,000 sentence test set that was provided to Polyglot Technology for evaluation.

BLEU Score Results for the TAUS Test Set



BLEU Score Results for Publicly Available Test Sets for the Language Pair

No widely shared public test sets are available for the language pair English-Swedish.

Description of Publicly Available Test Sets

wmt: News text test sets published by the Conference on Machine Translation¹

iwslt: Transcribed TED talk test sets published by the Conference on Spoken Language Translation²

Use Case Specific Evaluation

When employing machine translation for a specific use case, it is advisable to evaluate the systems with usage-scenario specific source text and its human reference translation. Maybe you have already data from a previous, similar project, or your translation vendor can help you create the test data. Polyglot Technology can assist in implementing a robust evaluation program.

¹ More details on the data sets can be found out e.g. for the 2020 edition on <http://statmt.org/wmt20/translation-task.html>

² Conference website: <https://iwslt.org/>

Interpreting BLEU Scores

The paragraphs in this section are adapted from Google AutoML Translate's [documentation page on evaluation](#) which is licensed under the [Creative Commons 4.0 Attribution License](#).

[BLEU \(BiLingual Evaluation Understudy\)](#) is a metric for automatically evaluating machine-translated text. The BLEU score is a number between zero and one that measures the similarity of the machine-translated text to a set of high quality reference translations. A value of 0 means that the machine-translated output has no overlap with the reference translation (low quality) while a value of 1 means there is perfect overlap with the reference translations (high quality).

It has been shown that BLEU scores correlate well with human judgment of translation quality. Note that even human translators do not achieve a perfect score of 1.0 (for the reason that a source sentence can have several valid, equally appropriate translations).

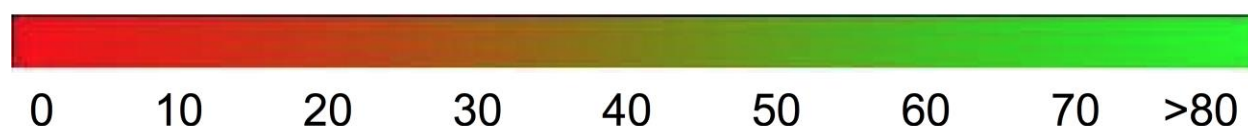
Interpretation

Trying to compare BLEU scores across different corpora and languages is strongly discouraged. Even comparing BLEU scores for the same corpus but with different numbers of reference translations can be highly misleading.

However, as a rough guideline, the following interpretation of BLEU scores (expressed as percentages rather than decimals) might be helpful.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

The following color gradient can be used as a general scale [interpretation of the BLEU score](#):



Segment Analysis

To determine which translations improved the most between the non-customized and the customized Amazon Translate we ranked the translations by the most improved [COMET](#) metric³.

The table below shows the 30 most improved translations with a visualization of edits that have to be applied to transform the Amazon Translate translation into the Amazon Translate Active Custom Translation.

source	reference	difference
Dispense Wash Fluid & Aspirate failed.	Dispense Wash Fluid & Aspirate misslyckades.	Dispense ring av tvättvätska och aspirering <u>Wash Fluid & Aspirate</u> misslyckades.
System is in the Warming Up, Cleaning, Diagnostics, Check Status, or Cleaning Failed state.	Systemet är i läget Warming Up, Cleaning, Diagnostics, Check Status eller Cleaning Failed.	Systemet är i läget Uppvärmning, Rengöring, Diagnostics, Check Status eller Rengöring misslyckades <u>Uppvärmning, Rengöring, Diagnostics, Check Status eller Cleaning Failed</u> .
This Readme contains important information about AutoCAD OEM 2012 that may not be included in Help.	Den här Viktigt-filen innehåller väsentlig information om AutoCAD OEM 2012 som kanske inte finns med i hjälpen.	Denna <u>Viktigt-fil</u> innehåller viktig information om AutoCAD OEM 2012 som kanske inte finns med <u>finns med</u> i hjälpen.
For purposes of calculating ROAE, a figurative fiscal	För att beräkna den genomsnittliga avkastningen på eget kapital har en	Vid beräkningen av avkastning på avkastning på avkastning på avkastning på <u>Vid beräkningen av avkastning på avkastning på avkastning på</u>

³ BLEU works only really well on a corpus (entire text) level than on an individual sentence level

source	reference	difference
effect ('effetto fiscale figurativo') has been taken into account, when determining net profit (see p. 163 of the initial business plan).	figurativ skatterättslig verkan (effetto fiscale figurativo) beaktats vid fastställandet av nettovinst (se s. 163 i den ursprungliga affärsplanen).	avkastning på avkastning på investeringen har man tagit hänsyn till en på eget kapital har en figurativ finanspolitisk effekt ("effetto fiscale figurativo") o beaktats vid fastställandet av nettovinsten (se s. 163 i den ursprungliga affärsplanen).
The total number of tests (primary reagent, vial) or milliliters (ancillary reagent) for all onboard packs is equal to or less than the Reagent Warning Counts you entered at the	Det totala antalet tester (primärreagens, ampull) eller milliliter (ancillary-reagens) för alla pack i instrumentet är lika med eller mindre än de värden som du har angett för Reagent Warning Counts i	Det totala antalet tester (primärreagens, flaskaampull) eller milliliter (ancillary-reagens) för alla inbyggda förpackningar pack i instrumentet är lika med eller mindre än de värden som du har angett för Reagensvt Warningkontroller som du angav på Counts i
Drafting tool set Open Shapes tool group (expanded) Point Style	Verktygsuppsättningen Drafting verktygsgruppen Open Shapes (utvidgad) Point Style	Verktygsuppsättning för ritn Draftings verktygsgruppen Öppna former (expanderutvidgad) pPunktformat
Cheeses made from sheep's milk or buffalo milk, in containers containing brine, or in sheepskin or goatskin bottles	Ost tillverkad uteslutande av fårmjolk eller buffelmjolk i behållare innehållande saltlake eller i behållare av får- eller getskinn	Ost framställd av fårmjolk eller buffelmjolk, i behållare innehållande saltlösningake eller i fårskinnbehållare av får- eller getskinn
stable at -20°C for 60 days after reconstitution.	stabil vid -20°C i 60 dagar efter rekonstituering.	sStabil vid -20°C i 60 dagar efter beredningrekonstitution.
TIMP-1 ELISA Controls Controls have been developed to provide customers with control material for quality monitoring of day-to-day assay performance.	TIMP-1 ELISA-kontrollerna togs fram som material för kvalitetskontroll för daglig övervakning av analysförfarandet.	TIMP-1-ELISA-C-kontroll Controls Controls Cler Kontrollslerna har utvecklats för att förse kunderna med kontrollmaterial för kvalitetsövervakning av de nt dagliga analysprestandanarbetet.
United States of America Federal Regulations.	Amerikas förenta staters federala föreskrifter.	Amerikas förenta stater Fs federal Regulationsa regler.
Training CD 2, Reaction Tube's Journey & Periodic Maintenance, is divided into two parts.	Övnings-CD 2, Reaction Tube 's Journey & Periodic Maintenance, är uppdelad i två delar.	UtbildTräning-s-CD 2, Reaktionsförens resa och p Tube's Journey & Periodiskt underhåll, är uppdelad i två delar.
The listener parameter can be a reference to either a function or a class method.	Parametern listener kan vara en referens till antingen en funktion eller en klassmetod.	AvlyssnarparameternParametern listener kan vara en referens till antingen en funktion eller en klassmetod.
Your M409WX Projecter includes all of the benefits of Dell's award winning service and support and Next Business Day Exchange Service for two years as standard.	Med M409WX-projektorn medföljer alla förmåner som ingår i Dells prisbelönta service och support samt 2 års ersättningsservice nästa arbetsdag som standard.	DinMed M409WX-projekter innehåller alla fördelarna medorn följer alla förmåner som ingår i Dells prisbelönta service och support och samt 2 års ersättningsservice nästa arbetsdag i två år som standard.

source	reference	difference
Certified Partner Case Studies> Dell and JMC PowerEdge case study> Nexus Open PowerEdge Systems> Dell and MWI PowerEdge systems	Fallstudier för certifierade partners> Fallstudie – PowerEdge från Dell och JMC> Nexus Open PowerEdge-system> PowerEdge-system från Dell och MWI	Fallstudier <u>för certifierade partners</u> > Dell och JMC PowerEdge-fallstudie > Nexus Open PowerEdge-system> Dell och MWI PowerEdge-system
Show option to select Package in Knowledge Center Resource Type drop down if 1. Set to 0 to hide.	Visa alternativ för att välja paket i kunskapscentrets nedrullningsmeny för resurstyp om 1. Ställ in som 0 för att dölja.	Visa alternativ för att välja Paket i <u>Knowledge Centerekunskapscentrets nedrullningsmeny för Resource Type listrutan</u> om 1. Ställ in till 0 för att dölja.
Components such as the ScrollPane and the UILoader have complete events that allow you to determine when content has finished loading.	Komponenter som ScrollPane och UILoader har complete-händelser som gör att du kan avgöra när innehållet har lästs in fullständigt.	Komponenter som ScrollPane och UILoader har <u>complete</u> händelser som gör att du kan avgöra när innehållet har lästs in.
"Dell M993s 19" Performance CRT Monitor - MG"	Dell M993s 19-tums Performance CRT-skärm – MG	"Dell M993s 19" <u>CRT-skärm med prestanda-tums Performance CRT-skärm</u> – MG"
The risk charge factor per each of the 20 EEA regions for Windstorm, taking into consideration the effect of diversification effect between zones.	Katastrofriskfaktorn för var och en av de 20 EES-regionerna för storm, med hänsyn tagen till diversifieringseffekten mellan områden.	<u>RiskdebiteringsKatastrofrisk</u> faktorn för var och en av de 20 EES-regionerna för <u>storm</u> storm, med <u>beaktande av effekterna av hänsyn tagen till</u> diversifiering <u>effekten</u> mellan <u>zonerområden</u> .
The operator should follow the recommendations to prevent the transmission of infectious agents in health-care settings as recommended for potentially infectious specimens in Protection of Laboratory Workers from Infectious Disease Transmitted by Blood, Body Fluids, and Tissue, 2nd edition; Approved Guideline (1997) Document M29-A, National Committee for Clinical Laboratory Standards (NCCLS).	Operatören bör följa rekommendationerna för att förhindra överföring av smittämnen i hälsovårdsmiljöer enligt bestämmelserna för potentiellt smittsamma prov i Protection of Laboratory Workers from Infectious Disease Transmitted by Blood, Body Fluids, and Tissue, andra upplagan; Approved Guideline (1997) dokument M29-A, National Committee for Clinical Laboratory Standards (NCCLS).	<u>Operatören bör följa rekommendationerna för att förhindra överföring av smittämnen i hälso-och sjukvårdsmiljöer som rekommenderas enligt bestämmelserna för potentiellt smittsamma prover i skydd av laboratoriearbetare från i Protection of Laboratory Workers from Infektionssjukdomar som överförs av blod, kroppsvätskor och vävnad, 2:us Disease Transmitted by Blood, Body Fluids, and Tissue, andra upplagan; gGodkänd riktlinje (1997) Dokument M29-A, National Cella kommittéen för Eklinical Lska laboratory Sjestandardser</u> (NCCLS).
This could be due to a bad TimeCycl.iml, a motor, or board(2) problem	Detta kan betyda fel på filen TimeCycl.iml, en motor eller problem med ett board (2).	Detta kan bero på ett dåligt <u>problem med tTimeCycl.iml</u> , en motor eller <u>bräda (2)ett board (2) problem</u>
Controls run immediately following an adjustment	Kontroller som körs omedelbart efter en justering bör ligga inom acceptabla gränser.	Kontroller <u>na som</u> körs omedelbart efter en justering bör ligga inom acceptabla gränser.

source	reference	difference
should be within acceptable limits.		
Will accommodate various types of Fiber interface trnasceivers (e.g. 1000SX, 1000LX; modules sold separately)	Passar för olika typer av fibergränssnittsmottagare (t.ex. 1000SX, 1000LX – moduler säljs separat)	Kommer att rymma Passar olika typer av fibergränssnitt-strnasceivers (t.ex. 1000SX, 1000LX; m. Moduler na säljs separat).
The serum and plasma values from 56 donors were determined for blood collected in serum and plasma (lithium heparin and EDTA) collection tubes.	Serum- och plasmavärden från 56 givare bestämdes på blod taget i serum- och plasmaprovtagningsrör (litiumheparin och EDTA).	Serum- och plasmavärdena från 56 donatorer bestämdes för på blod uppsamlat i serum- och plasma provtagningsrör (litiumheparin och EDTA) uppsamlingsrör.
The CV (coefficient of variation) for the low calibrator replicate RLU values for Reagent A was too large.	Variationskoefficienten (CV) för replikatet för den låga kalibrators RLU-värden för Reagens A är för hög.	Variationskoefficienten (variationskoefficienten) för RLU-värdena CV för replikatet för det låga kalibrator replikatets RLU-värden för Reagens A var för stor hög.
Detailed information on Support & Deployment Services.	Detaljerad information om tjänster för support och driftsättning.	Detaljerad information om Support & Deployment Services och driftsättningstjänster.
EUR 0,04 per barn layer chick per week for a maximum of 326450 chicks for a maximum amount of EUR 81743,18;	0,04 euro per värphönskyckling i stall per vecka för högst 326450 kycklingar; stödet får uppgå till högst 81743,18 euro,	0,04 euro per ladhönskyckling i stall per vecka för högst 326450 kycklingar till ett högsta belopp på; stödet får uppgå till högst 81743,18 euro.
It needs to verify whether the capacity created by the project is less than 5% of the size of the market measured using apparent consumption data of the product concerned, unless the average annual growth rate of its apparent consumption over the last five years is above the average annual growth rate of the EEA's GDP.	Det ska således kontrolleras huruvida den kapacitet som skapas genom investeringsprojektet motsvarar mindre än 5% av marknaden, mätt med hjälp av uppgifter om den synliga konsumtionen av produkten i fråga, om inte den genomsnittliga årliga tillväxttakten i den synliga konsumtionen under de senaste fem åren legat över den genomsnittliga årliga BNP-tillväxten inom EES.	D Kommissionen måste kontrollera om den kapacitet som skapas genom projektet är mindre än 5% av marknadens storlek mätt med hjälp av uppgifter om den synliga konsumtionen av den berörda produkten, såvida inte den genomsnittliga årliga tillväxttakten för dess synliga förbrukning under de senaste fem åren ligger över EES den genomsnittliga årliga tillväxttakt för EES." s BNP.
In order to ensure uniform conditions for the implementation of provisions on monitoring user compliance, detailed rules are required regarding the declarations to be made by recipients of funding for research involving the	För att säkerställa enhetliga villkor för genomförandet av bestämmelserna om övervakning av användarnas efterlevnad krävs tillämpningsföreskrifter för de deklarationer som ska lämnas av den som tar emot finansiering för forskning som inbegriper användning av genetiska resurser	För att säkerställa enhetliga villkor för genomförandet av bestämmelserna om övervakning av användarnas efterlevnad krävs närmare detaljerade bestämmelser för om de deklarationer som ska lämnas av mottagare av finansiering för forskning som inbegriper användning av genetiska resurser och traditionell kunskap som rör genetiska resurser. Resurser samt uppgifter om de

source	reference	difference
utilisation of genetic resources and traditional knowledge associated with genetic resources, as well as regarding the declarations to be made by users at the stage of final development of a product developed via the utilisation of genetic resources and traditional knowledge associated with genetic resources.	och traditionell kunskap som rör genetiska resurser, samt för de deklarationer som ska göras av användarna i slutskedet av produktutvecklingen för en produkt som utvecklats med hjälp av användning av genetiska resurser och traditionell kunskap som rör genetiska resurser.	deklarationer som användarna ska lämna i <u>under</u> slutliga utvecklingen av en produkt som utvecklats genom användning av genetiska resurser och traditionell kunskap som rör genetiska resurser.
Additional information may be provided by the manufacturers.	Tillverkarna får tillhandahålla ytterligare information.	Ytterligare information får lämnas av tillverkarna. Tillverkarna får tillhandahålla ytterligare information.
"And a "BODY" following the tree structure and classes of elements of any one of the bodies included in this annex under points (b) to (g) depending on the nature of the information to be exchanged automatically."	"samt en del ("BODY") som följer trädstrukturen och kategorierna av upplysningar i någon av de delar som ingår i denna bilaga under punkterna b till g, beroende på vilken typ av upplysningar som ska utbytas automatiskt."	"Och en "KROPP" "BODY" som följer trädstrukturen och klasserna av beståndsdelar i någon en av de organ som ä <u>inges</u> i denna bilaga under <u>leden</u> b—g beroende på arten av den vilken typ av information som ska ut utbytas automatiskt."

Results Summary

By customizing Amazon Translate Active Custom Translation with TAUS Medical/Pharma English-Swedish training data the test set BLEU score improves by more than 5 points. This is a significant improvement in machine translation quality for this language pair and domain.

This numerical score improvement is supported by analyzing segments that improved significantly according to the semantic quality measure COMET: translations became more fluent and adhered better to the terminology used in the Medical/Pharma domain.

Polyglot Technology

Polyglot Technology LLC helps customers succeed with machine translation by ensuring that they make best use of data available to them, by assessing machine translation quality independent from MT vendors and by advising customers on how to best integrate the technology with people and processes.