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

Language Pair

English → Norwegian

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

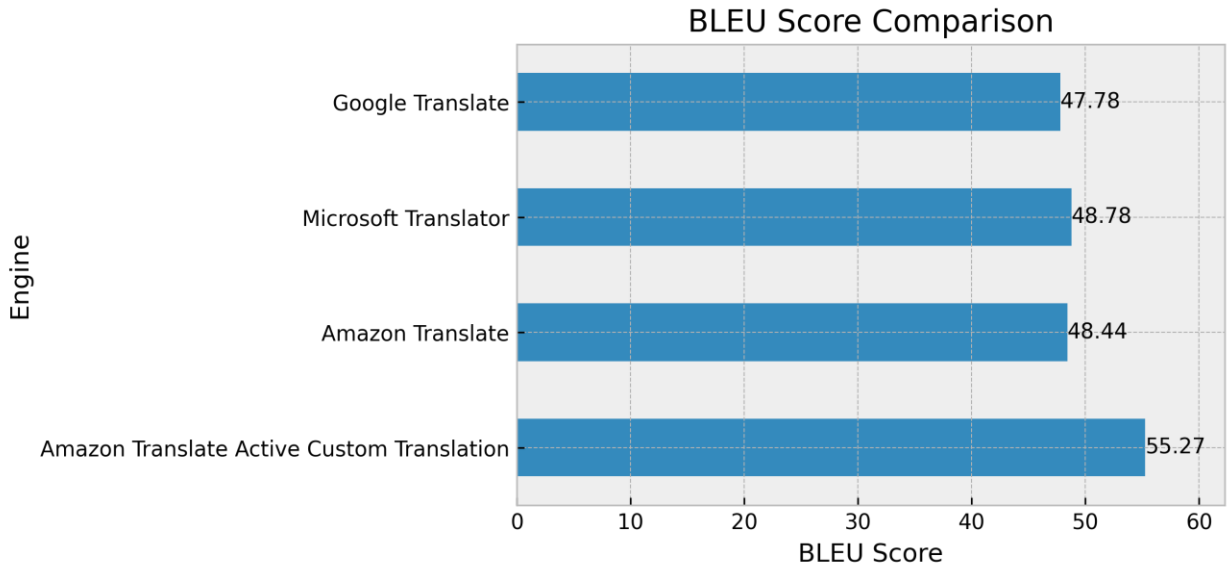
To judge whether machine translation is good or not, human evaluation is the best method. We can ask speakers of the source language and the target language, or better professional translators, to judge whether a machine translation is an adequate and fluent translation of the original text. Or we can 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-Norwegian.

Description of Publicly Available Test Sets

wmt: News text test sets published by the Conference on Machine Translation[1]

iwslt: Transcribed TED talk test sets published by the Conference on Spoken Language Translation[2]

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

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).

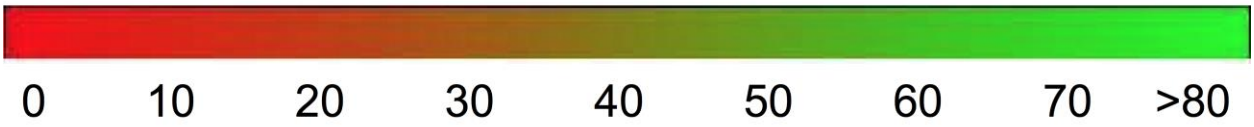
Interpretation

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 metric[3].

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
Support Package	Støttepakke	S~~upport P~~tøttepakke
Regulation difference	Reguleringsforskjell	Regulering~~-~~sforskjell
NOx Adsorber deSOx Regeneration Status	NOx-oppsamler deSOx regenereringsstatus	NOx ~~Adsorb~~-oppsamler ~~D~~deS~~o~~Ox ~~R~~regenerering ~~S~~sstatus
Preserve Infrastructure Investment	Utnytt infrastrukturinvesteringer	Bevar ~~Investment~~infrastruktur investering
Canada, United States.	Canada, USA	Canada~~, Amerikas forente stater~~ i USA.
Connection between transfer case control module and transfer case positioning motor	Forbindelse mellom transfercasens kontrollmodul og transfercasen posisjoneringsmotor	Forbindelse mellom ~~overføring tilfelle kontrollmodul og overføring tilfelle~~transfercasens kontrollmodul og transfercasens posisjonering~~-~~smotor

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

source	reference	difference
In Western Europe, spending on IT services, which includes IT planning, implementation, operations, maintenance and support, and IT training and education, will have a slightly higher CAGR of 4% for lower medium-size SMBs; upper medium-size SMBs will have a 3.9% CAGR for the period of 2009-2014.	I Vest-Europa vil investeringene i IT-tjenester, dvs. blant annet IT-planlegging, -implementering, -drift, -vedlikehold og -støtte samt IT-opplæring og -utdanning, oppleve en litt høyere årlig vekst (CAGR) på 4 % for virksomheter i nedre mellomklasse sammenlignet med 3,9 % for virksomheter i øvre mellomklasse i perioden 2009–2014.	I Vest-Europa vil utgifter til IT-tjenester, som inkluderer IT-planlegging, -implementering, -drift, -vedlikehold og -støtte, samt IT-opplæring og -utdanning, ha en noe høyere CAGR på 4% for små og mellomstore små og mellomstore mellomstore små og mellomstore små og mellomstore bedriftervirksomheter i nedre mellomklasse vil ha en 3,9% CAGR for perioden 2009-—2014.
Lunch Hour To < Lunch Hour From!	Lunsjtid til < Lunsjtid fra!	Lunsj Hour Ttime til < Lunsj Hour Ftime fra!
Composite RGB	Sammensatt RGB	KomposiSammensatt RGB
Emit CIDFontType2 As CIDFontType2 (PS Version 2015 And Greater)	Send CIDFontType2 som CIDFontType2 (PS-versjon 2015 og senere)	EitAvgi CidFonttyPEType2 som CidFonttyPEType2 (PS- versjon 2015 og større)
Identification numbers IV	Identifikasjonsnumre IV	iIdentifikasjons nummer IV
Client Activation Pack - Motivation	Aktiveringspakker for klienter – motivasjon	Klient Aktivering Pack - Mspakker for kunder — motiv
Special	Spesiell	Spesiaell
Rack Weight 35.94 lbs (23 Kg), maximum configuration	Vekt for rack er 23 kg (35,94 lbs), maksimal konfigurasjon	Vekt påfor rack er 35, 94 pund (23 kg), maksimal konfigurasjon
In Asia Pacifi c, spending on IT services, which includes IT planning, implementation, operations, maintenance and support, and IT training and education, will have a CAGR of 8.8% for upper medium-size SMBs compared with 7.3% CAGR for lower medium-size SMBs for the period of 2009-2014.	I Asia og Stillehavsregionen vil investeringene i IT-tjenester, dvs. blant annet IT-planlegging, -implementering, -drift, -vedlikehold og -støtte samt IT-opplæring og -utdanning, oppleve en årlig vekst (CAGR) på 8,8 % for virksomheter i øvre mellomklasse sammenlignet med 7,3 % for virksomheter i nedre mellomklasse i perioden 2009–2014.	I Asia Pacifi c vil utgifter til IT-tjenester, som inkluderer IT-planlegging, -implementering, -drift, -vedlikehold og -støtte, og samt IT-opplæring og -utdanning, ha en CAGR på 8,8% for mellomstore mellomstore små og mellomstore små og mellomstore bedrifter sammenlignet med 7,3% CAGR for små og mellomstore små og mellomstore bedrifter forvirksomheter i øvre mellomklasse sammenlignet med 7,3% CAGR for virksomheter i nedre mellomklasse i perioden 2009-—2014.
Make Primary	Gjør primær	Gjør Pprimær
Fuel tank level , plausibility, Fill level too high, (DFC_SCRPODPlausUTnkLvlLo)	Drivstofftank nivå, sannsynlighet, fyllenivå for høyt, (DFC_SCRPODPlausUTnkLvlLo)	Drivstofftank nivå, plausibilitsannsynlighet, Ffyll-enivå for høyt, (DFC_Scrpodp CRPODPlausUtTnklvllLo)
Triggering of the driver's air bag, passenger air bag disarmed	Utløsing av førerens airbag, passasjers airbag utkoblet	Utløsing av førerens kollisjonspute, passasjer kollisjonsputenairbag, passasjer airbag avvæpnet
If you suppress leading zeros, 0.5 becomes .5, and if you suppress	Hvis du undertrykker foranstilte nuller, blir 0.5 til .5, og hvis du	Hvis du undertrykker nuller somforanstilte nuller, blir 0,.5 .5, og hvis du undertrykker etterfølgende nuller, blir 0,.5000 0,5.

source	reference	difference
trailing zeros, 0.5000 becomes 0.5.	undertrykker etterstilte nuller blir 0.5000 til 0.5.	
Storage Data Reduction.	Reduksjon av datalagring.	LReduksjon av lagringsdatareduksjon.
The processor, or CPU, processes the critical information and instructions that make your Dell Vostro laptop perform.	Prosessoren, eller CPU-en, behandler den kritiske informasjonen og instruksjonene som får din Dell Vostro bærbar PC til å yte.	Prosessoren, eller prosessorCPU-en, behandler den kritiske informasjonen og instruksjonene som får den bærbarein Dell Vostro-PC enbærbar PC til å utføfungere.
Remotely heal systems with AMT technology even if systems are powered down or OS is hung (via Intel Management Engine or MEBX)	Reparer systemer eksternt med AMT-teknologi, selv om systemene er slått av eller operativsystemet henger (via Intel Management Engine eller MEBX)	Eksternt helbrede systemerReparer systemer eksternt med AMT-teknologi selv om systemene er slått av eller OS eroperativsystemet hengter (via Intel Management Engine eller MEBX)
Time to run the initiation is between 5-10 minutes.	Tid for å kjøre initieringen er mellom 5-10 minutter.	Tid tilfor å kjøre innvielsitieringen er mellom 5-10 minutter.
● After shutting down the high voltage system and removing the 12 V battery negative (-) terminal, wait at least three (3) minutes to discharge the air bag capacitor.	● Etter at høyspentsystemet er slått av og 12 V-batteriets negative (–) polsko er frakoblet, vent i minst tre (3) minutter for at kollisjonsputekondensatoren skal være utladet.	● Etter at du har slått av høyspenningstsystemet og fjerneter slått av og 12 V--batteriets negative (-) terminalpolsko er frakoblet, vent i minst tre (3) minutter for å tømmeat kollisjonsputen kondensatoren skal være utladet.
SCR inducement warning to the driver - Caused by Low Level ADBLUE, Second Alarm	SCR indusering advarsel til fører – forårsaket av lavt nivå ADBLUE, andre alarm	SCR inducementsering advarsel til sjåfører — forårsaket av lavt nivå ADBLUE, Second Aandre alarm
Retirement Risk	Pensjoneringsrisiko	Retirement RPensjonsrisiko
You can explode a compound object, such as a polyline, dimension, hatch, or block reference, to convert it into individual elements.	Du kan splitte opp et sammensatt objekt som en polylinje, dimensjon, skravering eller blokkreferanse for å konvertere det til enkeltstående elementer.	Du kan eksplodereitte opp et sammensatt objekt, for eksempel en polylinje, dimensjon, skravering eller blokkreferanse, for å konvertere det til enkeltstående elementer.
The Optical Drive selected cannot be quoted with Blood Orange Base, Battery, Spare Battery or LCD.	Den valgte optiske stasjonen kan ikke kombineres med Blood Orange-base, -batteri, -reservebatteri eller -LCD.	Den valgte optiske stasjonen kan ikke oppgikombineres med Blood Orange-Bbase, B-batteri, R-reservebatteri eller -LCD.
Immediate Shutdown Signal "A"	Umiddelbar avstegnings-signal "A"	Umiddelbar Shutdown Savslutningssignal «A»
Pressure Control Solenoid "L"	Trykkontrollsolenoid «L»	Trykkkontroll-Ssolenoid «L»

Results Summary

By customizing Amazon Translate Active Custom Translation with TAUS Medical/Pharma English-Norwegian training data the test set BLEU score improves by close to 7 points. This is an impressive 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 preferred 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.