DeMT™

TAUS DeMT™ Evaluation Report

June 2022

The landscape of online machine translation services is evolving quickly, providing ever more languages, integrations and customization options. To help customers navigate this landscape TAUS performs dedicated data and customization work to provide a machine translation service optimized for vertical domains and customer-specific implementations under the TAUS DeMT™ name.

This report provides an independent evaluation of this service benchmarked against the major online machine translation providers Amazon, Microsoft and Google. The first version of this report was published in December 2021. With changes in the machine translation services and the investment in more data and customization work it was time to update the report. In the future, even more frequent updates are planned to present the most up-to-date information.

Introduction

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

Businesses often want to translate text with different styles and topics. For enterprise use, online machine translation engines offer customization with existing translations that reflect the desired style and topic. This data is often called “parallel data”.

TAUS makes such customization data available via the TAUS Data Marketplace, and uses the data to offer the TAUS Data-enhanced Machine Translation (DeMT™) service.

TAUS asked Polyglot Technology LLC to independently benchmark the machine translation quality of DeMT™ relative to other major online machine translation providers.

Online machine translation systems get constantly updated. TAUS is keeping up with these changes by investing heavily in both data and customization work (described below) to provide customers with the best performing DeMT™ engines.

To continuously monitor these updates, we are planning to publish the TAUS DeMT™ Evaluation Report quarterly, adding additional online machine translation services and domains as needed.

TAUS Data and Customization for DeMT™

Data Work

Data for DeMT™ can come from different sources. The main source for the data sets in this report is the Matching Data library. The Matching Data library is a set of domain centered corpora, all originating from the language data repositories that TAUS has built over the years from human created translations. These corpora are generated from querying the repositories using domain-specific, real-life documents, which return data that matches the domain searched for.

In order to use a domain specific bilingual corpus as a set for training or customizing a translation model, further steps of cleaning and filtering are required. Basic filtering using language recognition, deduplication, segment length threshold, among others, is the first pass. Filtering on sentence embedding score is the second pass. The outcome of these filters is a selection of only the most reliable data.

Training Amazon Active Custom Translation, Google AutoML Translation and Microsoft Custom Translator all require a split between training data and testing data. The method here is a randomized split, where 2000 segments are set aside for test translations. This same set is used for each of the MT providers. Other requirements for the training of the models depend on the MT provider. Microsoft Custom Translator and Google AutoML Translation need a validation set, to be extracted from the training dataset.

Customization Work

Customization and training of models is a highly specialized work. Apart from the preparation of consistent, relevant and high-quality datasets, all MT providers have their own way of working. They differ widely in the way they interact with the user, the complexity and power of their API, their storage solutions, what can be exported and in what format, and even terminology in some cases is different. TAUS DeMT™ takes that discrepancy away, and runs a uniform workflow for all the providers, with uniform metrics to compare their performance.

Machine Translation Evaluation

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

The BLEU Score

Automatic metrics that also use human reference translations, have been developed to calculate a numeric score for machine translation quality. For 20 years now 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.

Most recently other scores like COMET have been developed to reflect human judgements of translation quality better than BLEU. We make COMET scores available in detailed reports that are available on request. For this report we use the familiar BLEU score which is recognized and understood widely. With our curated test set the large BLEU score differences (see below) accurately reflect the differences in machine translation quality between the evaluated systems.

To calculate BLEU for this report we use sacreBLEU[1]. We would like to thank Matt Post and all contributors to sacreBLEU for making this invaluable tool available.

TAUS Test Set

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

Summary

The source language for the evaluated systems is always English and the target languages are various European languages – in total we evaluated 8 language pairs for the Ecommerce domain, 18 language pairs for the Medical/Pharma domain and 4 language pairs for the Financial domain.

TAUS DeMT™ Translate improves the BLEU score over the *worst* performing non-customized online machine translation service:

by more than 10 points, or 25% on average
by more than 5 BLEU points at a minimum

Keep in mind that if you choose an online machine translation at random, or you choose the default online machine translation provider for your organization, you might choose the worst performing online machine translation service for a language pair.

Even if the best performing online machine translation service for the domain is known, TAUS DeMT™ Translate still provides a significant quality boost. TAUS DeMT™ Translate improves the BLEU score over the *best* performing non-customized online machine translation service:

by more than 5 points, or 11% on average
by more than 1 BLEU points at a minimum

[1] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

These improvements demonstrate the superiority of TAUS DeMT™ Translate for the Ecommerce, Medical/Pharma and Financial domains over non-customized online machine translation services.

See the following figures for BLEU scores on the TAUS test data - the charts are sorted language-wise by most to least improvement of TAUS DeMT™ Translate over the worst performing online machine translation service.

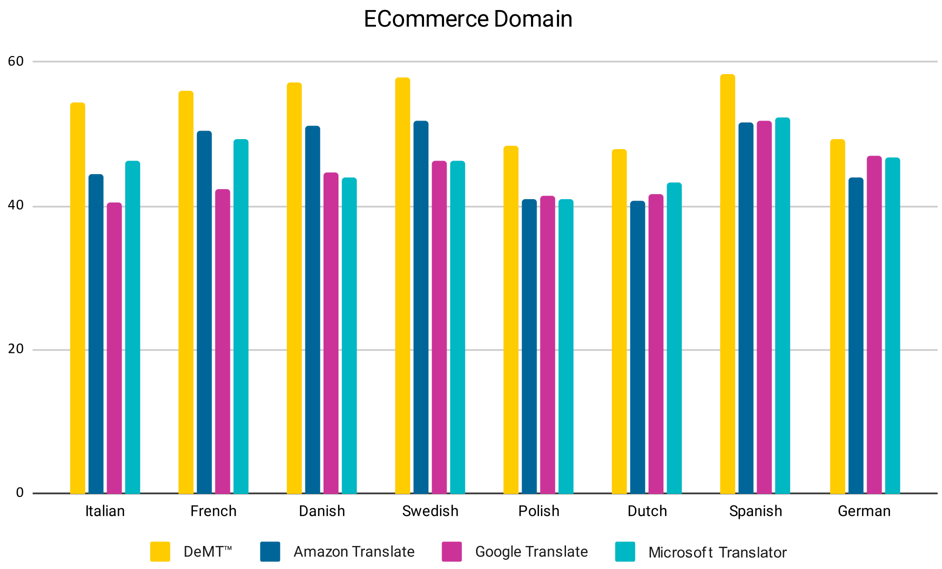


Figure 1: BLEU Scores for the TAUS Test Sets for the Ecommerce Domain

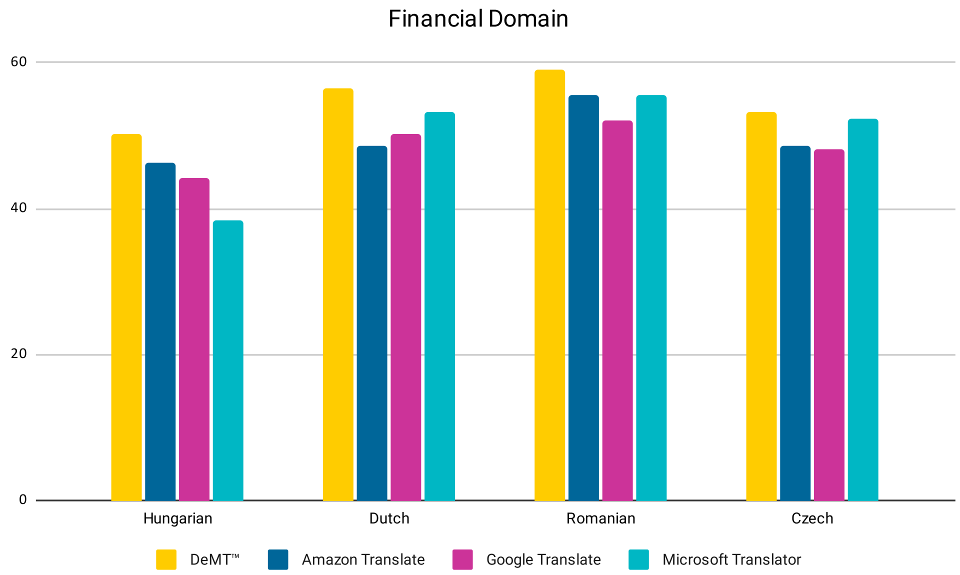


Figure 2: BLEU Scores for the TAUS Test Sets for the Financial Domain

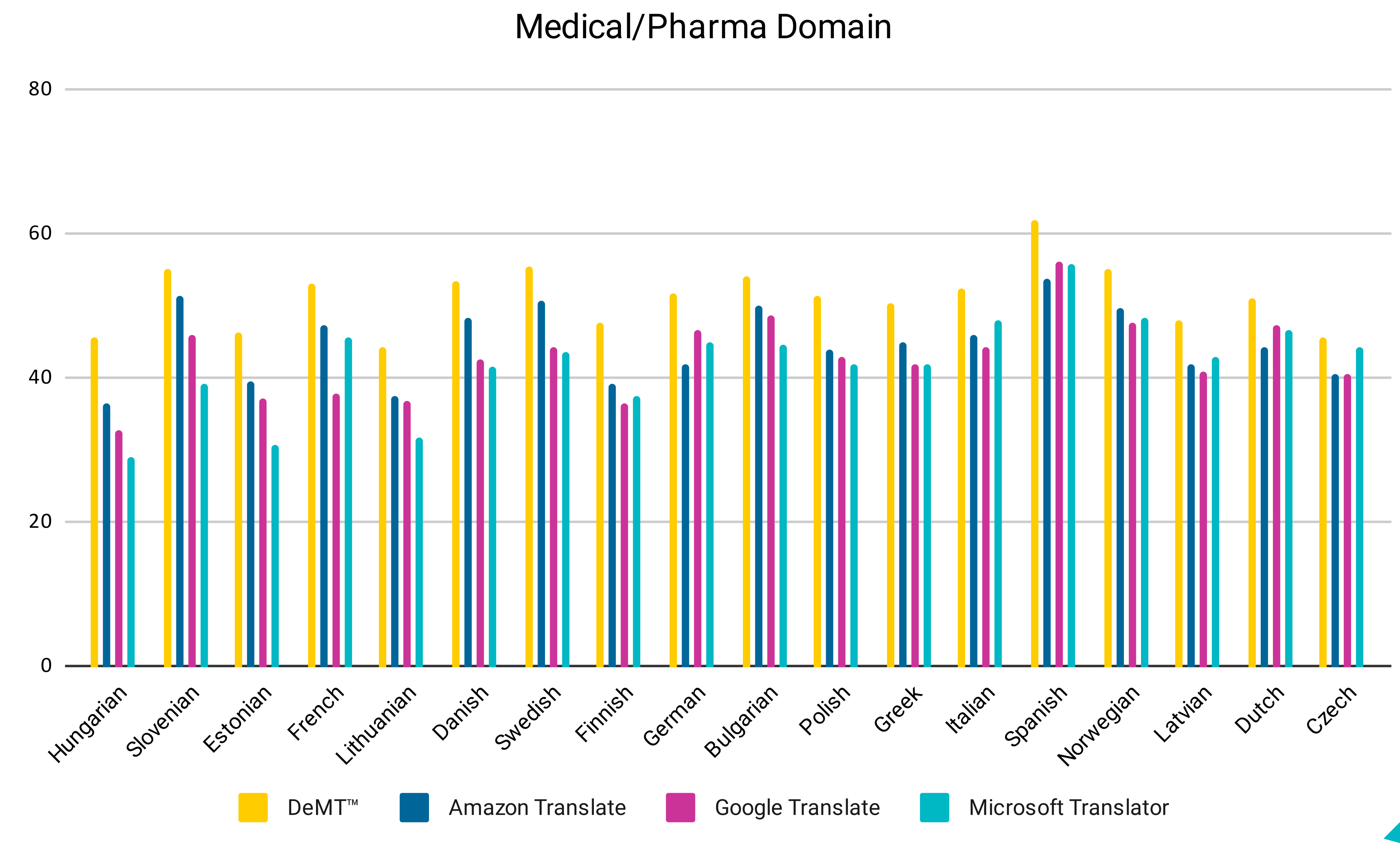


Figure 3: BLEU Scores for the TAUS Test Sets for the Medical/Pharma Domain

Use Case for Domain-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 human reference translations. Maybe you already have 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.

When you go through the effort of compiling use-case specific data it is likely worth it to consider obtaining personalized training data with the TAUS Matching Data service. This requires gathering use-case specific source text independent from the test data – a so-called “query set”. This query set can then be used to create highly specific training data using TAUS Matching Data. In many cases this can improve machine translation quality even more than pre-selected domain training data.

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).

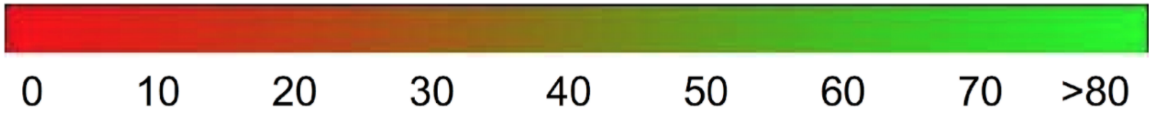
Interpretation

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 interpretation of BLEU scores on the next page (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:



About Polyglot Technology

Polyglot Technology LLC helps customers succeed with machine translation by enabling them to 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.

About TAUS DeMT™

TAUS DeMT™ aims at delivering the maximum achievable quality of MT output through a dedicated focus on language data and NLP technology. TAUS has the following offerings of DeMT™:

1. For companies that don't invest in their own translation technology and want to buy the best quality customized MT output as a service: **DeMT™ Translate**
2. For companies that customize and maintain their own translation technology and are in need of training data and NLP support to further improve the performance of their engines: **DeMT™ Build**
3. For companies that want to evaluate and benchmark the performance of their translation technology: **DeMT™ Evaluate** (in partnership with Polyglot Technology)
4. For enterprises that want to integrate a customized MT solution in their content and translation workflow solution: **DeMT™ Enterprise**

Contact TAUS for more information.