TAUS
the language data network

September 2023

TAUS DeMT™ Evaluation Report.

Introduction

Just over a year ago, in June 2022, TAUS published the DeMT™ Evaluation Report. Intended as a periodic check on the state of the art in customizable machine translation (MT), it is now about time to revisit the statistics and run the tests again.
Changes in online MT services happen fast. We seem to keep repeating that, but the way it changes is never as expected, and takes us by surprise each time. In the year since the last evaluation report, we have seen the rapid rise of the large language models (LLM). This had, and will continue to have, an impact on the translation industry as a whole. The avalanche of research publications in this field shows this potential, and various experiments explore the possibilities to learn, evaluate and annotate bilingual content using LLMs. The innovation does not stop there. New types of data collection and synthetic data generation offered by LLMs will find their way into the curation of training data. With that, training data is expected to be collected from a wider range of sources, and together with the improved machine learning capabilities, this will be key for further enhancement of MT in the coming time.

Fitting Context

This report builds on the work that was started with the DeMT™ Evaluation Report of June 2022. Back then, TAUS was exploring the ways in which translations of large online machine engines can be improved by training on domain-specific data. The idea is that generic MT has gotten to a level that is good for all-purpose translations, but it lacks the finesse to generate text that is properly fitting for its context. The translation is understandable at best, but still cumbersome to read and prone to misunderstanding.

The first aspect to consider in this context is domain-specific translation. We expect documents on medical subjects to communicate in an entirely different tone from what can be found on e-commerce websites for personal care. What is true for 'top-level' domains, such as finance, news, e-commerce, business, can also be said about specializations within those domains that have their own vocabularies, styles, or predominant content types.

In the medical domain, which is one of the domains included in this report, we can distinguish a few bigger subdomains. One subdomain, for example, can be communication about health-related issues, such as nutrition advice, articles regarding health issues, or doctor - patient conversations. This shows a big contrast with medical research, where the language is more academic, and usually describes scientific procedures, statistical analysis, and biomedical terminology. The third important subdomain within the medical domain is about rules and regulations, policy, and assurance. The subdomain here originates from organizations within the medical sphere, but shares much of its language with business, policy, and law.

Following this logic, it means that translation should not only reflect the general appropriateness to the domain, but, ideally, also the specific style and tone that fits the corporate communication. With the DeMT™ proposition, TAUS targets exactly that. Success of DeMT™ is driven by two components: data collection and the responsiveness of the MT engine.

Data Collection

TAUS collects data from the following sources: public repositories, human data creation, crawling, and synthetic data generation. For the purpose of MT training, more data used to be always better. However, now the quantity alone is getting less relevant as language models are getting more responsive to training. The tendency is towards fewer data, which is especially important if you want to move away from generic models: bigger buckets of training data inevitably create more generic models.

The newer iterations of online MT engines can work on fewer training data. That is in itself positive, but it also puts more significance on the quality of the data collection. With fewer data the tolerance for noise becomes even smaller.

For this evaluation report we chose to keep the same training and test datasets. We are aware that the point for customization could be made even clearer by further refinement of the training sets applying the latest methods of data curation. At TAUS, data collection and curation is under continuous development: focus on a narrower subdomain, with additional cleaning steps to further optimize the customization. However, selecting such an 'in-lab' approach would also lose applicability for real-world situations where MT users often deal with somewhat fuzzy data. Also, comparability with the previous report was important in this decision. We were very curious to explore how the engines evolved.

Responsiveness of machine translation

As this report shows, the major MT engines are performing better at customization. The evaluation includes Microsoft Custom Translator, and Amazon Active Custom Translation.We ran comparisons of the customized versions against the baseline engines of both providers. Microsoft Custom Translator has recently been updated with a new version. End-users will not immediately see this, but it shows in the improved customization. Also Amazon got better scores for their customized engine.

The previous DeMT™ evaluation report showed quite some differences between the engines in terms of the responsiveness to training. Spending your resources on training and getting hardly any result can be a frustrating experience. Striking the right balance between domain relevance, quality and quantity is already a process of trial and error that takes experience to get acquainted with. All the effort can go to waste when the engine has a low responsiveness to training.

Another problem that keeps customized translation from improving significantly can occur when it overfits the training. In that case customization seems to follow each quirk from the training set, which might result in a big difference from the baseline translation, but not necessarily an improvement. A lot of the translations are changed by the customized training, but on segment level, the changes seem rather random.

Evaluation

The big online MT engines tend to illustrate their customization improvements exclusively in BLEU scores. Because of the ubiquity of BLEU scores, the number feels familiar, and most people in the translation industry know how to interpret the meaning of a BLEU score. That said, BLEU as a metric has important flaws - it is dependent on reference sentences, and does not compensate for valid synonyms or morphological variants, to name a few of the problems.

The previous report used only BLEU scores for its analysis. For comparability, we will still report on BLEU scores, but two metrics are added to compensate for the shortcomings of BLEU: chrF and COMET scores.

We calculate the BLEU scores according to the sacreBLEU library, using the standard parameters. chrF scores use reference translations as well, but are known to have a good correlation with human evaluation. We use the library from https://github.com/m-popovic/chrF for the calculation, with all standard parameters. Finally, COMET scores take word embeddings into account for its evaluation, and therefore have a different approach to translation evaluation. We use COMET22-DA, the latest library described in https://github.com/Unbabel/COMET.

General Results

This report focuses on the test translations of 30 language pair and domain combinations. The language pairs all have a European language as target language and English as the source language. The domains are financial (4 language pairs), medical/pharmaceutical (18 language pairs), and e-commerce (8 language pairs). We used the same language/domain combinations as in the June 2022 report, and we worked with the same training and test sets. Despite the limitations of the BLEU score, we still took the opportunity to compare the scores of the previous report to the current version, in order to get an impression of the latest developments. These are the key observations:

1. Between June 2022 and September 2023, **the average BLEU score of the uncustomized test translations improved** for both the Microsoft and Amazon translation engines. No substantial improvements, though. Amazon's average baseline score in both our tests stayed between 46 and 47 BLEU points. Microsoft's average baseline score increased, around 1.5 points, making it slightly above 48 points on average. What is more interesting in the Microsoft BLEU scores, is that in the latest tests **there are less outliers on the lower end, thus showing a more uniform scoring for all languages.** In this way it is catching up with the scores of Amazon, which were comparatively more consistent in June 2022.

2. A bigger change in scores between the current and the previous report was achieved for the customized engines. As the charts illustrate, both Amazon Active Custom Translation and Microsoft Custom Translator averaged around 53 - 54 BLEU points. Compared to the previous report, Amazon improved this by 1.5 points, and Microsoft by almost 3 points. It's important to consider that these are averages over a large range of domain and language combinations, and it means that for each combination that scores above average, there are also lower scoring combinations. But on the positive side: the variety of **the improvements decreased, meaning that the improvement in BLEU score by customization is more probable and more uniform.** Whereas in the June 2022 report there were a few cases with a lower BLEU score after training, **now all the customizations show improvement**, almost always more than 3 points higher than the baseline model.

This report concludes with charts that summarize the test results. As mentioned, apart from BLEU scores, both the chrF and the COMET system scores were calculated. The general trend of the BLEU scores is largely confirmed by the other metrics, but they also add more detail for at least some of the domain-language combinations. The chrF scores are very similar to the BLEU scores. Where BLEU scores show a strong increase, the chrF scores do likewise, and vice-versa. The differences tend to be slightly less pronounced for chrF scores, but the trend is consistent.

COMET scores often have a much smaller range of values. Compared to the BLEU scores, the increasing scores of training seem somewhat dampened. In very rare cases this results in showing a negative response to training, whereas the other metrics still indicate a minor growth.The opposite also can be observed, where, mainly in the medical/pharmaceutical domain, for some languages the training of Microsoft Translator shows a major increase in COMET scores.

As the charts show, domain-specific MT customization can be very beneficial. The exact impact of training customized models, by its nature, remains somewhat unpredictable and that will probably not change any time soon. But what this report shows, is that **the customization itself has become more mature, and the results are getting more consistent**. In the time between our evaluations, the leading MT engines showed an overall progress and made it clear that customization in particular is paramount for an effective content translation strategy.