LILT

2021

Measuring and Comparing Machine Translation Quality

Learn about machine translation, the different ways it’s measured, and see how Lilt’s Adaptive Machine Translation quality compares to Google’s and Microsoft’s machine translation systems.

What is Machine Translation?

At its core, machine translation is fully automated software that translates content from one language to another.

It can be used to assist professional translators, where artificial intelligence (AI) and human intelligence are combined to produce the highest quality translations for critical content, often at a reduced cost compared to unassisted human translations. When used as a standalone technology, while not a substitute for a professional human linguist when translation requires a high quality guarantee, machine translation has made large amounts of content in the world accessible to people outside of the target language audience by providing fast and cost-effective translation into more languages.

MT systems are sometimes applied in localization for huge collections of documents (i.e., millions of words) in which none of the documents are going to be read many times. In those situations, traditional human translation wouldn't be feasible due to the sheer volume of content, so we turn to AI.

Types of Machine Translation Systems

There are multiple types of machine translation systems and different approaches to machine translation, most notably:

Rule-Based Machine Translation (RbMT)

Translation rules are created based on the grammar, syntax, and semantics of language. Linguists write down large sets of rules for each language pair (i.e., EN-ES, EN-FR, etc.). Content is then fed through these algorithms and translated into the appropriate language.

Statistical Machine Translation (SMT)

This approach automatically mines parallel text (pairs of source and target sentence-by-sentence translations) for translational equivalent fragments, words, or phrases that are statistically most likely to be appropriate for use. These phrases can then be recombined using a decoding algorithm.

These systems are able to be trained faster than RbMT and, more importantly, can be trained with much less human effort.

Neural Machine Translation (NMT)

While both statistical and neural MT use huge datasets of translated sentences to teach software to find the best translation, the models themselves are different. Unlike SMT, Neural MT uses neural networks to consider whole sentences when predicting translations, which allows it to take into account the context in which each word and phrase is used.

How are MT Systems Measured?

While machine translation has helped revolutionize the way languages are translated from a source language to a target language, it's important to understand exactly how to measure the success of machine translation output. There are a few common ways to measure the quality of a translation:

Round-Trip Translation

This type of measurement essentially takes a back and forth approach to checking quality. The source language is first translated into a target language, as it normally would be. Next, that output is re-translated back into the source language, thus completing its "round trip".

While easy, this approach tells you almost nothing about translation quality. It's a good party trick, perhaps, but systems that don't make a perfect round trip of your favorite movie quote are still often very useful for high-volume translation jobs.

Human Evaluation

Human experts in specific languages mainly check on two factors to judge the quality of a translation - adequacy and fluency. While this strategy is ideal in a lot of cases, not every translation can be sent to an expert panel, as it's time consuming and cost prohibitive, espeically as companies look to translate quickly. Non-expert human evaluation often misses important features of translation quality, so human evaluation is only effective if done with care and with the right evaluators.

Automatic Evaluation with Bilingual Evaluation Understudy (BLEU)

The BLEU score has long been the standard for evaluating fully automated MT. It compares a machine translation output with that of a human translation. The close the MT output is to a human translation, the higher the score.

To calculate an automatic score, you need one or more human translations for reference, then test the MT output against those reference points. BLEU scores are used frequently because they correlate reasonably well with human judgments of quality when comparing similar types of MT systems, and as a result, there is a lot of data to compare to.

Next-Word Prediction Accuracy (WPA)

When an interactive MT system is being used to assist a professional linguist, BLEU is not an appropriate measurement of translation quality. Instead, an effective evaluation metric is next-word prediction accurary. As the translation types, the system guesses what word they will type next, over and over. The proportion of the time that the system guesses correctly what the translator was going to type is the WPA.

Evaluating Adaptive Neural MT Using BLEU Scores.

To help our customers better understand the improvements that Lilt's Adaptive NMT model shows over Google's MT and Microsoft's MT, we decided to compare BLEU scores.

Lilt's model is adaptive, meaning that it can be customized to a particular kind of text by providing relevant example translations. This customization involves automatically adjusting the millions of numerical parameters of the system's neural network so that it becomes better at translating the provided examples and any similar text. This strategy is called Domain Adaptation, and is known to yield the best quality output for a specific domain.

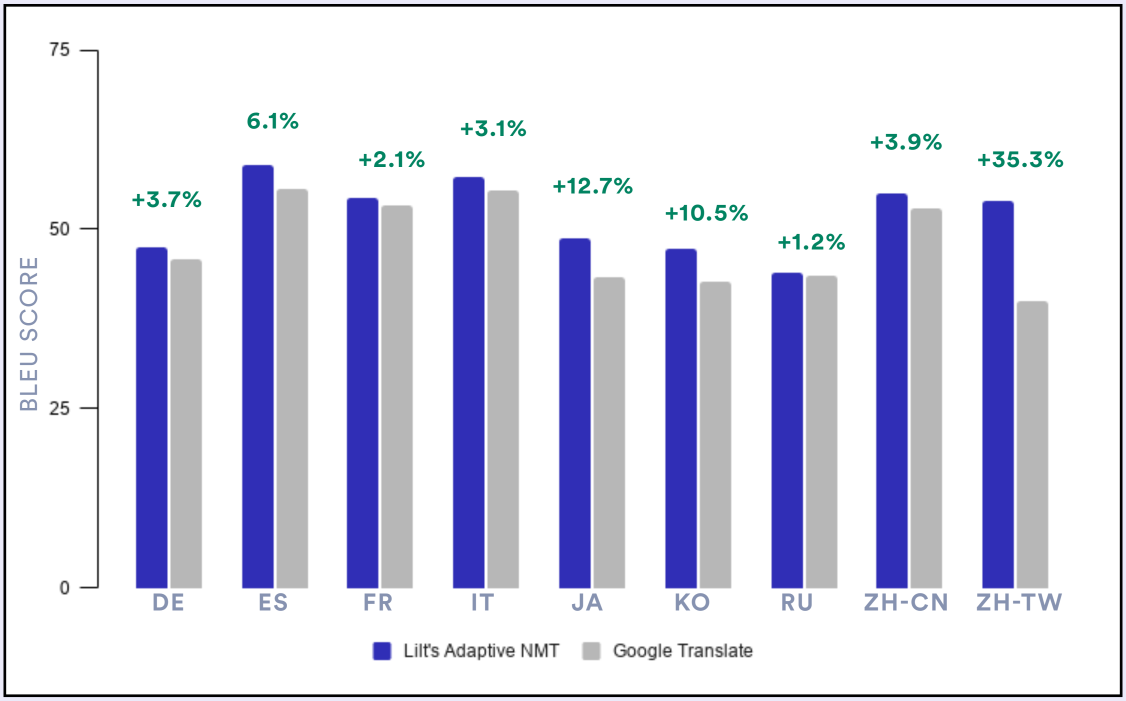
We wanted to compare two types of content from different domains - support content and marketing content. While support content tends to be more technical by nature (describing products and features), marketing content is more freeform and brand focus and thus has less repetitiveness in the content.

Once the system was trained, we looked at how well the machine translation output matched a reference human translation. We then compared the score to that produced by Google Translate (for marketing content) and Microsoft (for support content).

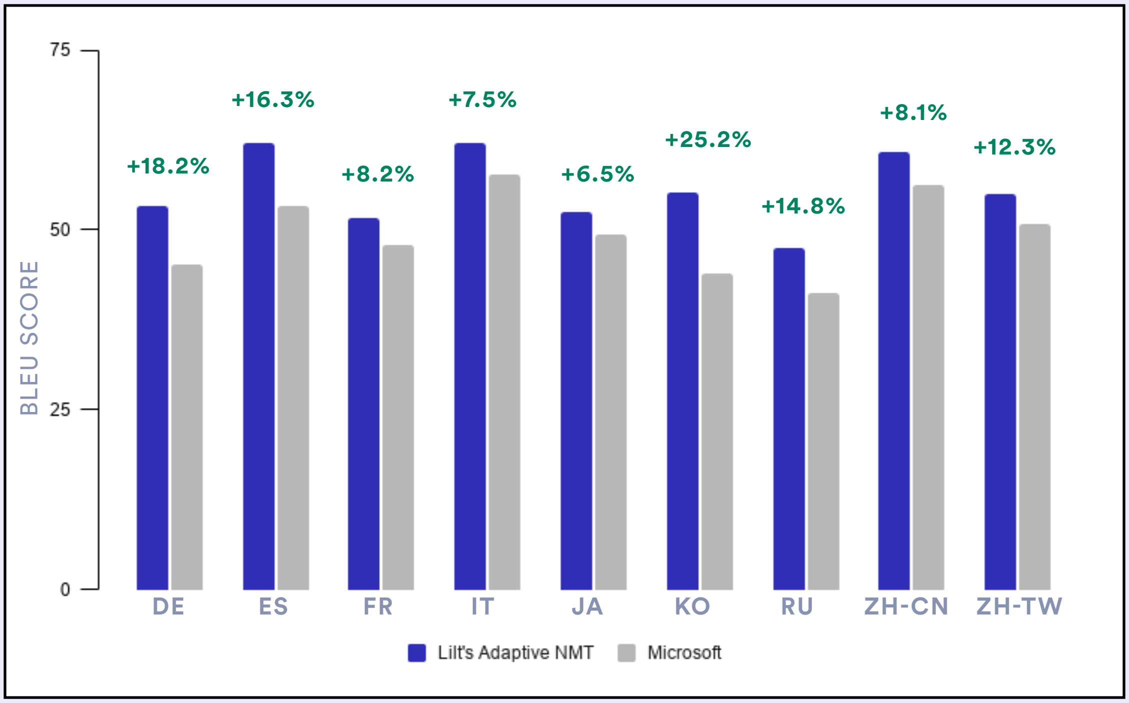
The experiment was run on a handful of language pairs, from English to a number of other languages. The Lilt system was adapted using translation memories containing up to 4.5M source words in a given language.

Comparing the Results.

MARKETING CONTENT



SUPPORT CONTENT



Deciphering the Data.

As shown in the data above, Lilt's adaptation technique was able to outperform Google's MT system by an average of 7.7% and outperform Microsoft's MT system by an average of 12.3%.

For certain languages, the difference in performance was small - marketing content in French only performed 2.06% better. However, for Korean support content, Lilt's Adaptive NMT performed 25.23% better than Microsoft's MT.

What does this mean overall? Adaptive machine translation systems have large advantages over the world's best non-adaptive neural systems. The ability to customize the machine translation engine to a company's domain makes an adaptive solution the clear winner for many users. While a company can curate its own data, that can be error prone and time consuming, as systems and data have to be configured and managed manually.

With a model like Lilt, however, the system constantly makes translation suggestions to the translator as they work through the content. As the translator accepts or rejects those suggestions, the system automatically and continuously adapts to the work and improves over time. The results get better without any configuration or system management, improving the translation quality and allowing all involved to be more efficient.

Looking Beyond Machine Translation Quality.

Ultimately, raw machine translation output and quality is only one part of the whole translation process.

For many, machine translation quality, while good, will not suffice depending on the type of content needed. While quality is important, it's just as important (maybe more) to have a skilled translator in the loop to ensure quality throughout the entire translation workflow.

Using an adaptive neural machine translation model like Lilt makes it even easier for the translator to provide feedback and make the MT system even better over time. Instead of strictly relying on linguists to edit machine translation output, they can be involved throughout, accepting or rejecting translation suggestions. Once they do, the model takes that input and learns over time, improving suggestions and increasing the overall quality. That process, known as human-in-the-loop, is helping to revolutionize the way translations are completed.

We're looking forward to researching and measuring the improvements that human-in-the-loop systems provide overall. Look out for additional whitepapers, blog posts, and webinars covering that data and more at www.lilt.com!

Choose a Strategic Partner

From engineering, product, marketing, sales, and support – there are many factors to consider when expanding to a new locale. One of the most critical aspects is to partner with a localization team dedicated to helping you succeed in measurable ways.

When choosing translation software and translators, partner with a company that provides a thorough proof of concept (POC) service during the sales process. You can measure both the quality of translations as well as efficiency gains. Through this process, you can better estimate the time-to-market and cost of localization.

Visit lilt.com for more information on our human-in-the-loop technology, and to see how we can help your business reach new markets – faster than ever before.

LILT