Omniscien™
TECHNOLOGIES

A SHORT GUIDE TO COMPARING MACHINE TRANSLATION ENGINES

"What is your BLEU score?" This is the single most ***irrelevant question*** relating to translation quality, yet one of the most frequently asked. BLEU scores and other translation quality metrics greatly depend on many factors that must be understood in order for a score to be meaningful. A BLEU score of 20 in some cases can be better than a BLEU score of 50 or vice versa. Without understanding how a test set was measured and other details such as language pair and domain complexity, a BLEU score is nothing more than a meaningless number.

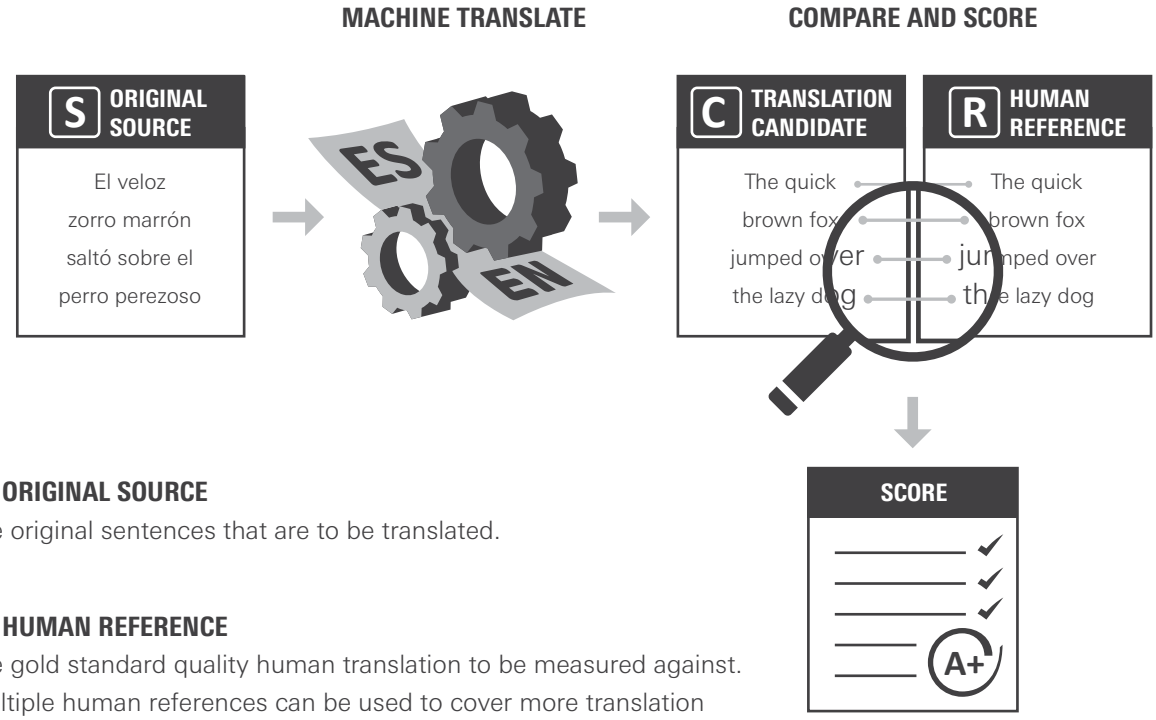
BLEU scores and other translation quality metrics will vary based upon:

- **THE TEST SET BEING MEASURED:** Different test sets will give very different scores. A test set that is out of domain will usually score lower than a test set that is in the domain of the translation engine being tested. The quality of the test set should be gold standard. Lower quality test set data will give a less meaningful score.
- **HOW MANY HUMAN REFERENCE TRANSLATIONS WERE USED:** If there is more than one human reference translation, the resulting BLEU score will be higher as there are more opportunities for the machine translation to match part of the reference.
- **THE COMPLEXITY OF THE LANGUAGE PAIR:** Spanish is a simpler language in terms of grammar and structure than Finnish or Chinese. Typically, if the source or target language is more complex then the BLEU score will be lower.

- **THE COMPLEXITY OF THE DOMAIN:** A patent has far more complex text and structure than a children's story book. Very different metric scores will be calculated based on the complexity of the domain. It is not practical to compare two different test sets and conclude that one translation engine is better than the other.
- **THE CAPITALIZATION OF THE SEGMENTS BEING MEASURED:** When comparing metrics, the most common form of measurement is Case Insensitive. However when publishing, Case Sensitive is also important and may also be measured.
- **THE MEASUREMENT SOFTWARE:** There are many measurement tools for translation quality. Each may vary slightly with respect to how a score is calculated, or the settings for the measurement tools may not be set the same. The same measurement software should be used for all measurements. Omniscien Technologies provides a tool free of charge that measures a variety of quality metrics.

It is clear from the above list of variations that a BLEU score number in isolation without these variables clearly defined has no real meaning.

HOW BLEU SCORES AND OTHER TRANSLATION METRICS ARE MEASURED



S ORIGINAL SOURCE

- The original sentences that are to be translated.

R HUMAN REFERENCE

- The gold standard quality human translation to be measured against.
- Multiple human references can be used to cover more translation variations.

C TRANSLATION CANDIDATE

- The translated output from the machine translation engine being measured.
- Multiple translation candidates can be measured at one time. This is useful to compare multiple translation engines against each other e.g. Omniscien Technologies, Google and Systran or different versions of the same translation engine.

With BLEU scores, a higher score indicates higher quality. A BLEU score is not a linear metric. A 2 BLEU point increase from 20 to 22 will be considerably more noticeable than the same increase from 50 to 52. F-Measure and METEOR also work in this manner where a higher score is also better. For Translation Error Rate (TER), a lower score is a better score. Language Studio™ provides a tool that supports all of these metrics and can be downloaded for free.

BASIC TEST SET CRITERIA

The criteria specified by this checklist are absolute. Not complying with any of the checklist items will result in a score that is unreliable and less meaningful.

- **TEST SET DATA SHOULD BE VERY HIGH QUALITY:** If the test set data are of low quality, then the metric delivered cannot be relied upon. Automatically selecting a test set from translation memories does not guarantee quality. Test sets should have human quality review.
- **TEST SET SHOULD BE IN DOMAIN:** The test set should represent the type of information that you are going to translate. The domain, writing style and vocabulary should be representative of what you intend to translate. Testing on out-of-domain text will not result in a useful metric.
- **TEST SET DATA MUST NOT BE INCLUDED IN THE TRAINING DATA:** If you are creating an SMT engine, then you must make sure that the data you are testing with or very similar data are not in the data that the engine was trained with. If the test data are in the training data the scores will be artificially high and will not represent the same level of quality that will be in the output when other data are translated.
- **TEST SET DATA SHOULD BE DATA THAT CAN BE TRANSLATED:** Test set segments should have a minimal amount of dates, times, numbers and names. While a valid part of a segment, they are not parts of the segment that are translated; they are usually transformed or mapped. A focus for a test set should be on words that are to be translated.
- **TEST SET DATA SHOULD HAVE SEGMENTS THAT ARE BETWEEN 8 AND 15 WORDS IN LENGTH:** Short segments will artificially raise the quality scores as most metrics do not take into account segment length. Short segments are more likely to get a perfect match of the entire phrase, which is not a translation and is more like a 100% match with a translation memory. The longer the segment, the more opportunity there is for variations on what is being translated. This will result in artificially lower scores, even if the translation is good. A small number of segments shorter than 8 words or longer than 15 words are acceptable, but these should be very few.
- **TEST SET SHOULD BE AT LEAST 1,000 SEGMENTS:** While it is possible to get a metric from shorter test sets, a reasonable statistic representation of the metric can only be created when there are sufficient segments to build statistics from. When there are only a low number of segments, small anomalies in one or two segments can raise or reduce the test set score artificially.

COMPARING TRANSLATION ENGINES - INITIAL ASSESSMENT CHECKLIST

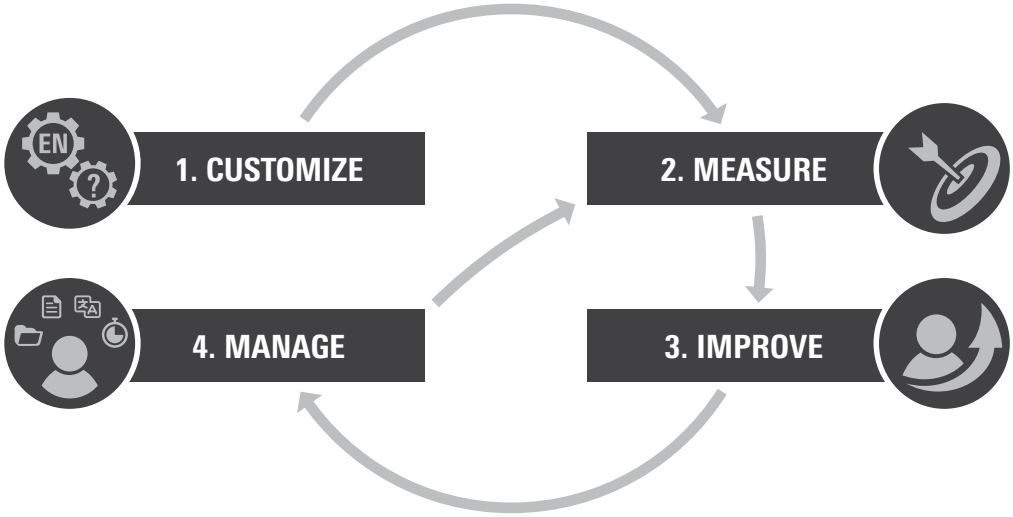
Language Studio™ can be used for calculating BLEU, TER, F-Measure and METEOR scores.

- **ALL CONDITIONS OF THE BASIC TEST SET CRITERIA MUST BE MET:** If any condition is not met, then the results of the test could be flawed and not meaningful or reliable.
- **TEST SET MUST BE CONSISTENT:** The exact same test set must be used for comparison across all translation engines. Do not use different test sets for different engines.
- **TEST SETS SHOULD BE "BLIND":** If the MT engine has seen the test set before or included the test set data in the training data, then the quality of the output will be artificially high and not represent a true metric.
- **TEST SET MUST BE CARRIED OUT TRANSPARENTLY:** Where possible, submit the data yourself to the MT engine and get it back immediately. Do not rely on a third party to submit the data. If there are no tools or APIs for test

set submission, the test set should be returned within 10 minutes of being submitted to the vendor via email. This removes any possibility of the MT vendor tampering with the output or fine tuning the engine based on the output

- **WORD SEGMENTATION AND TOKENIZATION MUST BE CONSISTENT:** If Word Segmentation is required (i.e. for languages such as Chinese, Japanese and Thai) then the same word segmentation tool should be used on the reference translations and all the machine translation outputs. The same tokenization should also be used. Language Studio™ provides a simple means to ensure all tokenization is consistent with its embedded tokenization technology.

ABILITY TO IMPROVE IS MORE IMPORTANT THAN INITIAL TRANSLATION ENGINE QUALITY



The initial scores of a machine translation engine, while indicative of quality, should be viewed as a starting point for rapid improvement which is measured by the test set and BLEU scores. Depending on the volume and quality of data provided to the SMT vendor for learning from, the quality may be lower or higher. **More important than the initial quality is how quickly translation engine quality will improve.**

Frequently a new translation engine will have gaps in vocabulary and grammatical coverage. **Other machine translation vendors' engines do not improve at all or merely improve very little unless huge volumes of data are added to the initial training data.** Most vendors recommend retraining once you have gathered a volume of additional data that is at least 20% of the size of the initial data that the engine was trained on. Even when this volume of data is added, only a small improvement is achieved. As a result, very few translation engines evolve in quality much further than their initial quality.

In stark contrast, Language Studio™ translation engines are created with millions of sentences of data that Omniscien Technologies has prepared in addition to the data that the customer provides. The translation engines improve rapidly with a very small amount of feedback. It is not uncommon to get a 1-2 BLEU score improvement with as little as a few thousand post-edited sentences. Language Studio™ has a unique 4 step approach that leverages the benefits of Clean Data SMT and manufactures additional learning data by directly analyzing the edits made to the machine translated output.

Consequently, only a small amount of post-edited feedback can improve Language Studio™ translation engine quality quite considerably, and it can do so at speeds much faster and with far less effort than with other machine translation vendors. *Omniscien Technologies provides complimentary Incremental Improvement Trainings to encourage rapid translation engine quality improvement with every full customization and also offers additional complimentary Incremental Improvement Trainings when word packages are purchased, greatly reducing Total Cost of Ownership (TCO).*

An investment in quality at the development stages of a translation engine impacts and reduces the cost of post editing directly while increasing post editing productivity. While some rules, normalization, glossary and non-translatable term work will assist in the speed of improvement, the fastest and most efficient way to improve Language Studio™ engines is to post edit the translations and feed them back into Language Studio™ for processing. The edits will be analyzed and new training data will be generated, directly addressing the primary cause of most errors. In other words, just post editing as part of a normal project will result in an immediate improvement. Little or no other extra effort is needed. By leveraging the standard post editing process, the effort and cost of improvement as well as the volume of data required in order to improve is greatly reduced.

Depending on the initial training data provided by the client, a small number of Incremental Improvement Trainings are usually sufficient for most Language Studio™ translation engines to improve to a quality level approaching near-human quality.

Other machine translation vendors are now also claiming to build systems based on Clean Data SMT. Closer investigation reveals that their definition of "cleaning" is not the same as Omniscien Technologies'. Removing formatting tags is not cleaning data. Language Studio™ analyzes translation memories and other training data and ensures that only the highest quality in domain data from trusted sources is included in the creation of your custom engine. The result is that improvements are rapid. Even with just a few thousand segments edited, the improvements are notable. When combined with Language Studio™ hybrid rules and an SMT approach to machine translation the quality of the translation output can increase by as much as 10, 20 or even 30 BLEU points between versions.

COMPARING TRANSLATION ENGINES – TRANSLATION QUALITY IMPROVEMENT ASSESSMENT

- **COMPARING VERSIONS:** When comparing improvements between versions of a translation engine from a single vendor, it is possible to work with just one test set, but the vendor must ensure that the test set remains "blind" and that the scores are not biased towards the test set. Only then can a meaningful representation of quality improvement be achieved.
- **COMPARING MACHINE TRANSLATION VENDORS:** When comparing translation engine output from different vendors, a second "blind" test set is often needed to measure improvement. While you can use the first test set, it is often difficult to ensure that the vendor did not adapt its system to better suit and be biased towards the test set and in doing so delivering an artificially high score. It is also possible for the proof read test set data to be added to engines training data which will also bias the score.

As a general rule, if you cannot be 100% certain that the vendor has not included the first test set data or adapted the engine to suit the test set, then a second "blind" test set is required. When a second test set is used, a measurement should be taken from the original translation engine and compared to the improved translation engine to give a meaningful result that can be trusted and relied upon.

BRINGING IT ALL TOGETHER

The table below shows a real world example of a version 1 translation engine from Omniscien Technologies and an improved version after feedback. Additional rules were added to the translation to meet specific requirements of the client, which resulted in considerable improvement in translation quality. This is part of Omniscien Technologies' standard customization process. Language Studio™ puts a very high level of control in the customer's hands where rules, runtime glossaries, non-translatable terms and other customization features ensure the quality of the output is as close to human quality and requires the least amount of editing possible.

BLEU Score Comparisons	Omniscien Technologies			Google	Bing	Systran
	V1 SMT	V2 SMT	V2 SMT+Rules			
Case Sensitive						
Reference 1	36.05	45.96	56.59	30.58	29.64	21.01
Reference 2	35.80	39.31	48.85	32.05	29.94	22.56
Reference 3	38.65	52.31	65.03	35.51	33.17	24.68
Combined References	**50.45**	**66.52**	**80.48**	**44.58**	**41.65**	**30.26**
Case Insensitive						
Reference 1	41.30	52.65	59.25	32.18	31.49	22.49
Reference 2	41.01	45.32	51.24	33.67	31.64	23.88
Reference 3	43.99	58.97	67.49	37.15	35.01	25.92
Combined References	**56.83**	**74.35**	**82.89**	**46.26**	**43.68**	**31.68**

*** Language Pair : English into French**
Domain : Information Technology

It can be seen clearly from the scores above that when all three human reference translations are combined the BLEU score is significantly higher and that the BLEU scores vary considerably between each of the human reference translations. The impact of the improvement and the application of client specific rules can also be seen, raising the case sensitive BLEU score from 50.45 to 80.48 (an increase of 30.03 in just one improvement iteration). One interesting side effect of having multiple human references is that it is often possible to judge the quality of the human reference also. In the example above, the machine translation output is much closer to human reference 3, indicating a higher quality reference. The client later confirmed that the editor who prepared the reference was a senior editor and more skilled than the other 2 editors who prepared human reference 1 and 2.

A BLEU score, as with other translation metrics, is just a meaningless number unless it is established in a controlled environment. Asking "What is your BLEU score?" could result in any one of the above scores being given. When controls are applied, translation metrics can be used both to measure improvements in a translation engine and to compare translation engines from different vendors. However, while automated metrics are useful, the ultimate measurement is still a human metric. Language Studio™ also provides tools to assist in delivering balanced and meaningful metrics for human quality assessment.

ABOUT OMNISCIEN TECHNOLOGIES

Omniscien Technologies is a leading global supplier of high- performance and secure high-quality Language Processing, Machine Translation (MT) and Machine Learning technologies and services for content intensive applications. Our wide range of solutions serves clientele from various industries including the Localization Industry, Online Research Services, Publishing, eCommerce, Media, Online Travel, Technology, Enterprise and Government.

Omniscien Technologies has gained a reputation for cutting edge solutions with its Language Studio™ platform. Depending upon the customer's unique requirements, Language Studio™ can be deployed in a variety of ways to integrate with their in-house data processing and translation management systems, and it offers unparalleled levels of customization and control as well as feature rich pre- and post- processing, enabling customers with even the most complex data to achieve both high quality and high volume output to satisfy every use case. Omniscien Technologies has by far the most comprehensive and feature rich system in the market today.

Covering 550 language pairs and with a number of industry specific solutions, Omniscien Technologies remains the partner of choice for customers with complex, high-volume bespoke data processing and machine translation needs.
Further Information

For further information on Omniscien Technologies or Language Studio™, please visit www.omniscien.com or contact sales@omniscien.com

Omniscien™ TECHNOLOGIES

www.omniscien.com
sales@omniscien.com | +66 (2) 381 5869