CoTERM: A Consistency-Oriented Term Metric for MT System Evaluation

Amir Hazem, Kyo Kageura
LISLab, Graduate School of Education.
The University of Tokyo.
amir.hazem@gmail.com, kyo@p.u-tokyo.ac.jp

Abstract

Proper treatment of terms is an important and critical aspect in machine translation. It is therefore necessary to use appropriate metrics to evaluate MT system outputs from terminology perspective. However, despite the great improvements witnessed in the recent NMT and LLM models, MT system evaluation metrics that shed light on specific aspects of term translations are yet to be fully explored. In this paper, we propose CoTERM, a new metric for automatic evaluation of term translations based on the Herfindahl-Hirshman Index (HHI). CoTERM measures target term closeness to one or more reference translations, taking into account the fundamental criteria for translating terms, i.e. (i) accuracy; (ii) consistency at document or corpus levels; and (iii) appropriateness to the domain conventions with regard to term variations. The proposed metric correlates strongly with human raters, and empirical evaluations of a wide range of NMTs and LLMs show that the best MT systems in standard metrics are not necessarily the best at treating terms. CoTERM is thus shown to be highly useful for diagnosing MT systems' term translation performance and conveniently seen as complementary to generic measures for MT system evaluations.

Keywords: Term Translation, MT, Evaluation Metrics, Term Consistency

1. Introduction

Common evaluation metrics for MT systems and MT tasks are essential for understanding and diagnosing the performance of NMTs and LLMs[1]. Generic evaluation metrics such as BLEU, BertScore or COMET have contributed greatly to the development of MT systems. As the performance of MT systems has improved, more fine-grained metrics are needed to evaluate specific aspects of translations.

Treatment of technical terms is one such aspect. Inconsistent use or mistranslation of terms may cause serious problems (Ambrožič et al., 2010; Karwacka, 2014). The importance of terms is well recognized in the translation industry. The MQM error typology widely used in translation industries lists term translations, including proper use of terms with respect to reference terminologies, as one of the top-level error types (Lommel et al., 2014)[2].

In the evaluation of MT systems, shared tasks focusing on terminology translations have been carried out recently and metrics for evaluating term translations have also been proposed (ibn Alam et al., 2021b; Semenov and Bojar, 2022; Semenov et al., 2023). While this work covers important aspects of term translations, there is room for further elaboration, e.g. the use of reference terminologies and diagnosis of term translation consistencies.

Against this backdrop, we propose CoTERM, a metric and a framework to evaluate MT systems on their ability to translate terms. CoTERM is based on Herfindahl-Hirshman Index (HHI) (Herfindahl, 1950; Hirschman, 1945), a widely-used market concentration measure, and covers essential criteria for term translations: (i) accuracy; (ii) consistency at the document or corpus levels for correct and incorrect translations; and (iii) appropriateness to domain conventions with regard to the flexibility and rigidity of term variations.

Table 1 gives examples that CoTERM is to typically evaluate. The contributions of this work include: 1) new metric and a methodology for term translation evaluations; 2) an empirical evaluation of term translations of a wide-range of MT models in English-French, English-Japanese and Japanese-English; 3) a Github[3] to facilitate the reproduction of our results and the application of CoTERM to other datasets, and 4) a web-based interface for human raters to make terminology-focused annotations and evaluations of MT system outputs.

2. Related Work

We recently witnessed a growing interest in the proper treatment of terms in MT research (Crego et al., 2016; Post et al., 2019; Semenov et al., 2023; Kim et al., 2024). In this situation, metrics that reflect requirements for treating terms are essential.

The standard generic MT evaluation metric is BLEU (Papineni et al., 2002). Other metrics were also proposed. The Word Error Rate (WER), originally used in speech recognition (Woodard and

[1] Henceforth "MT systems" for simplicity.

[2] See MQM typology at: https://themqm.org/

[3] https://github.com/termview/CoTERM

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8639–8661
11–16 May 2026.

Type	Example
Src (En)	On Wednesday, the World Health Organization (**WHO**) declared the **ongoing outbreak** of **COVID-19** — the disease caused by **coronavirus SARS-CoV-2** — to be a **pandemic**.
Tgt (Fr)	Mercredi, l'Organisation mondiale de la santé (**OMS**) a qualifié l'**épidémie en cours** de **COVID-19** (la maladie provoquée par le **coronavirus SARS-CoV-2**) de **pandémie**.
Transf	Mercredi, l'Organisation mondiale de la Santé (**OMS**) a déclaré que l'**éclosion actuelle** de **COVID-19** de la maladie causée par le **coronavirus** CoV-2 SRAS) était une **pandémie**.
Src (En)	By converting the **phase** detected by such a method to **angle** of projected **sheet beam**, there can be obtained **three-dimensional** shapes of the **object** in **real time**.
Tgt (Ja)	こうして, 検出された位相を投影シートビームの角度に変換することにより, 対象物の**3**次元形状が実時間で得られる。
Gemma2	その方法で検出された相を投影されたシートビームの角度に変換することで、リアルタイムで対象物の**3**次元形状を得ることができる。

Table 1: Examples in English-French and English-Japanese. Type: Src and Tgt respectively represent the source and its target reference. MT: Transf (Tiedemann, 2020) and Gemma2 (Team, 2024). Correct translations are shown in green, acceptable translations in blue, and wrong translations in red.

Nelson, 1982; Morris et al., 2004), computes the edit distance between the reference and the predicted sentences. METEOR (Banerjee and Lavie, 2005) is based on the harmonic mean of unigram precision and recall. More recently, semantic similarity measures such as BertScore (Zhang et al., 2020) and COMET (Rei et al., 2020) were proposed, which showed strong correlation with human translations. Kocmi and Federmann (2023b) introduced GEMBA: a GPT-based evaluation metric that uses prompting techniques to detect translation errors without the need for reference translations.

Several metrics for evaluating term translations by MT systems were introduced. ibn Alam et al. (2021a) adopted Exact Match (EM) accuracy, Window Overlap (WO), and Terminology-biased translation Edit Rate (TER_M). EM is a standard aggregate accuracy over each term in the test set. To take into account the place of terms, WO takes a window of n (often two or three) tokens around the term for both hypothesis and reference, and calculates the percentage of common tokens. Terminology-biased translation Edit Rate (TER_M) uses the edit distance (Snover et al., 2006) and penalizes errors of term tokens. Abe (2023) is in line with ibn Alam et al. (2021a). Semenov and Bojar (2022) introduced a metric that evaluates term consistency. It is defined based on the lists of translated term candidates and of pseudo-reference terms. Various metrics can be applied to the lists. They proposed a metric based on the percentage of correct term occurrences for each source term (we will refer to this metric as "TC"). It was used in a shared MT task with terminologies (Semenov et al., 2023).

In terminology and translation studies, it is established that the treatment of terms needs to satisfy the following criteria: (i) proper terms should be used (accuracy);(ii) the same terms should be used consistently within a document or a set of documents (consistency)[4]; and (iii) a limited range of variations is allowed to facilitate smooth communication (Felber, 1984; Rogers, 1997). In human translation, reference terminologies are essential with regard to the criteria (i) and (ii). In evaluating MTs, Semenov and Bojar (2022) addresses consistency, i.e. the criterion (ii), basically under an assumption that there is a single correct term translation for each term.

To evaluate the consistency of terms in human translations, Itagaki et al. (2007) introduced a metric called Consistency Index based on the Herfindahl-Hirshman Index (HHI), a widely-used market concentration measure (Herfindahl, 1950; Hirschman, 1945; Rhoades, 1993), and evaluated the consistency of terms in an English-Japanese corpus. Gašpar et al. (2022) also applied the HHI-based metric to measure the consistency of terms in Croatian–English, Latin–English and Latin–Croatian corpora of legal documents. These measures take into account the existence of more than one correct translation. As of now, this approach has not been adopted to evaluate term translations by MT systems. CoTERM is defined based on HHI. The formal definition of HHI and how CoTERM incorporated HHI are given in the next section.

[4]The MQM typology (https://themqm.org/) gives two types of inconsistencies, i.e. inconsistency with a given glossary and inconsistency within a document. The former corresponds to accuracy here.

3. Proposed Method

We propose CoTERM, an HHI-based metric (Itagaki et al., 2007) and a methodological framework to evaluate MT systems' term translation performance. CoTERM incorporates the three criteria (i)-(iii) for treating terms given in Section 2. While these criteria can be evaluated separately, it is important to define a unified metric, because these criteria are interrelated in the MT setups.

CoTERM assumes reference translations, which is common in automatic MT evaluation. Reference list of terms are also assumed, either as bilingual or multilingual terminologies, following the established procedure in human translation, or as annotations in the corpus. This differs from Semenov and Bojar (2022), which aims at making the process as automatic as possible. We provide scripts that facilitate the use of terminologies in the application of CoTERM. Also, CoTERM is based on HHI while the consistency measure proposed in Semenov and Bojar (2022) is based on the percentage of correct occurrences for each term and uses F1 score in Semenov et al. (2023).

3.1. HHI and the Consistency of Terms

Herfindahl-Hirshman Index (HHI) (Herfindahl, 1950; Hirschman, 1945) is a market concentration measure (Rhoades, 1993), defined as:

$$\text{HHI} = \sum_{i=1}^{n} (\text{MS}_i)^2 \quad (1)$$

where MS_i represents the market share percentage of firm i and n represents the number of firms competing in the market. HHI takes values from 0 (perfect competition) to 10,000 (perfect monopoly).

Using an English-Japanese parallel corpus of software user interface (UI) strings, Itagaki et al. (2007) adapted HHI to evaluate the consistency of translations of terms across 104 products within a corpus of 300K sentences. They calculated the consistency of the translation of terms such as *Web server* and *Value type* in products including *Visual Studio*, *.NET Framework*, *SQL Server*, *Outlook*, etc.

The market share (MS_i) of a firm i in HHI becomes the translation share of a translation i of a source term t, and n becomes the total number of different translations of t. Therefore, n different translations are competing for gaining the share of the translations of the occurrences of t. To evaluate the consistency of translations of a given term t across p products, Itagaki et al. (2007) defined Consistency Index (C_t) as follows:

$$C_t = \frac{1}{p} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{f_i}{k_j} \times 100\right)^2 \quad (2)$$

where each translation share i in a product j is calculated as the ratio of its frequency f_i to the total translation occurrences k_j ($= \sum_{i=1}^{n} f_i$) in j. Gašpar et al. (2022) extended C_t and proposed a mean average consistency over a set of 100 terms[5].

Itagaki et al. (2007) concluded that the consistency index can be effectively used for quality assurance of translation data. When the consistency is high, it can also be used as an indicator of translation stability. Hence, high consistency score should save time for translators as it indicates to simply follow existing translations identified in old data. Low consistency however, indicates that translators should not choose one of existing translation variations, on the contrary, they should carefully examine the context in which the term is used.

3.2. HHI and MT

Both Itagaki et al. (2007) and Gašpar et al. (2022) applied HHI to human translations, in which all translated terms were assumed to be correct. To evaluate MT systems, on the other hand, it is essential to take into account incorrect term translations.

C_t, however, tends to overestimate term consistency when incorrect translations exist. Let us consider an extreme example where the English term 'outbreak' has two French correct translations, i.e. 'épidémie' and 'épidémique'. Suppose that 'outbreak' occurs 88 times in a given corpus, and that an MT system made the correct translations 'épidémie' 29 times and 'épidémique' 1 time. The system also made wrong translations 58 times. C_t score, as straightforwardly applied, is then $C_{\text{outbreak}} = \left(\frac{29}{30} \times 100\right)^2 + \left(\frac{1}{30} \times 100\right)^2 = 9,355$ or 93.35[6], while $\text{Accuracy}_{\text{outbreak}} = \left(\frac{29+1}{88} \times 100\right) = 34.09\%$. The C_t score shows that the consistency of the MT system is very high (close to perfect monopoly of 'épidémie'), while its accuracy is low. In this situation, it is misleading to conclude that the MT system performs well in term translations based on C_t. In order to evaluate the quality of term translations of MT systems, we have to take into account incorrect term translations, i.e. accuracy.

3.3. Elements of CoTERM Metric

These observations motivated us to develop CoTERM[7], a metric for evaluating MT systems' performance of term translations that reflects the criteria for treating terms. Here, section 3.3.1 describes how CoTERM incorporates accuracy in the C_t index. Section 3.3.2 explains how our measure addresses consistency at the document and corpus levels. Section 3.3.3 elaborates on appropriateness of the domain as to the flexibility of term variations. Finally,

[5] Itagaki et al. (2007) normalized C_t from 0 to 100 while Gašpar et al. (2022) normalized C_t from 0 to 10.

[6] When the score is normalized to range from 0 to 100.

[7] https://github.com/termview/CoTERM

section 3.3.4 introduces negative CoTERM which incorporates consistency of translation errors. We adopt the annotation CoTERM$^+$ when calculating the consistency of correct translations while we use CoTERM$^-$ for when dealing with the consistency of wrong translations.

3.3.1. Accuracy

We incorporate accuracy to the Consistency Index (C_t) within the MT evaluation setup by making simple modifications to equation (2): 1) we set the number of products p to one, as it does not immediately apply to accuracy in MT evaluation scenario; and 2) we replace k by $|t|$, the number of occurrences of the source term t, which is interpreted as the total number of translations of t to be found in the MT result. The CoTERM$_t^+$ score for a source term t to be translated is computed as follows:

$$\text{CoTERM}_t^+ = \sum_{i=1}^{n} \left(\frac{f_i}{|t|} \times 100\right)^2 \quad (3)$$

where f_i is the frequency of the correct translation i of term t, n is the total number of correct translations of t found by the MT system, and k is replaced by $|t|$ to include the missing correct translations.

How $|t|$ incorporates accuracy to original C_t can be shown by using Example 1, which illustrates: 1) three English source sentences (S1, S2, S3); 2) their corresponding reference translations in French (R1, R2, R3); and 3) the MT translation candidate outputs (C1, C2, C3). The source terms are marked in bold. The correct translations are underlined and the wrong translations are strikethrough.

Given the term **spread** which appears 3 times in S1, S2 and S3 ($|t| = |\text{spread}| = 3$). We observe that it has been correctly translated by *propagation* in C1, and by *diffusion* in C2. It was however wrongly translated by *dissemination*[8] in C3. As C_t addresses consistency of correct translations only, the factor n, which is the set of correct translations, will be composed of *propagation* and *diffusion* and their total frequency is $k = 2$. By computing C_t: $C_{\text{spread}} = \left(\frac{1}{2} \times 100\right)^2 + \left(\frac{1}{2} \times 100\right)^2 = 5,000$ or 50[9]. But this does not tell anything about the missed translation of spread in the third sentence. By taking into account accuracy, we obtain a new score reflected in CoTERM$^+$ as follows: $\text{CoTERM}^+_{spread} = (\frac{1}{3} \times 100)^2 + (\frac{1}{3} \times 100)^2 = 2222,22$ or 22,22[9]. In this case, CoTERM$^+$ score reflects both consistency of correctly translated terms and the accuracy by including the wrong translation *dissemination* in the score computation. In line with Gašpar et al. (2022) and in order to evaluate MT systems over a list of source terms (T) to translate, the overall CoTERM$_T^+$ is given by:

$$\text{CoTERM}_T^+ = \frac{1}{T} \sum_{t=1}^{T} \text{CoTERM}_t^+ \quad (4)$$

where T is the number of different source terms t to translate. In Example 1, $T = 6$, the overall $C_T = 75$ and the overall CoTERM$_T^+ = 70.37$.

Example 1:

S1. The **spread** of **covid-19** keeps increasing.

R1. La propagation du covid-19 ne cesse d'augmenter.

C1. La propagation du covid-19 continue d'augmenter.

S2. The **WHO** supports the **spread** of **clean technologies**.

R2. L'OMS soutient la diffusion de technologies propres.

C2. L'OMS encouragent la diffusion de technologies propres.

S3. The **WHO** has declared that the **ongoing outbreak** of **covid-19** is a **pandemic** with high **spread**.

R3. L'OMS a declaré que l'épidémie en cours du **covid-19** est une pandémie avec un risque élevé de propagation rapide.

C3. L'OMS a déclaré que ~~l'éclosion en cours~~ du covid-19 était une pandémie avec une forte ~~dissemination~~.

3.3.2. Consistency

When evaluating MT systems, it is important to consider consistency at different levels. Itagaki et al. (2007), for instance, addressed consistency at the product level. To provide opportunities for finer-grained analyses, we generalize this concept and propose a parametrized metric for which the evaluator can define the segment that fits the purpose of the study. The segment can be a paragraph, a document, a set of documents, the entire corpus or any meaningful range of data. The metric, CoTERM$^+$, is based on the computation of CoTERM$_T^+$ over a set of segments (S) as follows:

$$\text{CoTERM}^+ = \frac{1}{S} \sum_{s=1}^{S} \left[\frac{1}{T_s} \sum_{t=1}^{T_s} \sum_{i=1}^{n_s} \left(\frac{f_i^t}{|t|} \times 100\right)^2 \right] \quad (5)$$

where T_s is the number of source terms within the segment s and n_s is the number of its corresponding correct translations found by the MT system. If $S = 1$, equations (4) and (5) are equivalent and the test set is treated as a single document.

[8] We assume that *dissemination* is a wrong translation in this explanatory context.

[9] When the value is normalized between [0, 100].

3.3.3. Appropriateness

Terms are by essence bound to a specific domain, and their treatment in translations should follow rules and conventions of the domain[10]. In principle, the same form of terms should be used consistently, but in some cases several target terms can be accepted as correct translations. Linguistic phenomena including morpho-syntactic variations such as partial abbreviations, derivations or inflections should be taken into account as well (Fernández-Silva and Kerremans, 2011; Freixa, 2022).

To cover these phenomena, we define the appropriateness of term translations that reflects the degree of rigidity or flexibility in term translations accepted in context and in domain. In a rigid evaluation scenario, we limit the admitted reference translations to a restricted list of terms defined by relevant experts. This often corresponds to a one-to-one reference translation pairs. In a flexible scenario, the list of acceptable terms is extended based on relevant criteria to include such variations as synonyms, term inflections, etc. For instance, the English term 'infected' can have several possible translations in French, i.e. the noun 'infection', the singular verb 'infecté', the feminine verb 'infectée', etc. The rigid scenario limits the correct term only to, e.g., 'infecté', while the flexible scenario takes into account possible variants that can be considered acceptable depending on the occurrences. CoTERM$^+$ in the rigid scenario, i.e. CoTERM^+_r, is then given by setting i in equation (5) to one as:

$$\text{CoTERM}^+_r = \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(\frac{f^t}{|t|}\times 100\right)^2\right] \quad (6)$$

CoTERM$^+$ in the flexible scenario, i.e. CoTERM^+_f, is equivalent to equation (5).

3.3.4. Negative CoTERM

In evaluating and diagnosing the performance of MT systems, it is essential to take into account patterns of errors. In order to diagnose the weakness of MT systems, we observe whether the erroneous terms are generated consistently or randomly. Assuming a segment being one document (S=1), we define negative CoTERM, which measures the consistency of errors, as:

$$\text{CoTERM}^- = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{f^t_i}{|t|}\times 100\right)^2 \quad (7)$$

where T is the number of source terms and n is the number of its erroneous translations.

[10]The French term *dissemination* is a possible translation of *spread*, but in Example 1, it is a wrong translation in the context of the pandemic.

Example 2:

S. The **WHO** has declared that the **ongoing outbreak** of **covid-19** is a **pandemic**.

R. **[**L'OMS a declaré que**]** l'épidémie **[**en cours du **covid-19]** est une pandémie.

C. **[**L'OMS a déclaré que**]** ~~l'éclosion~~ **[**~~en cours~~ du covid-19**]** était une pandémie.

The detection of erroneous translations is illustrated in Example 2 where the term **ongoing outbreak** in S, was wrongly translated by ~~l'éclosion en cours~~ in C. Using our heuristic algorithm, the wrong translation can be identified based on the longest common chunks between the reference (R) and the MT system prediction (C). The chunks in red: "L'OMS a déclaré que" and "du covid-19 était" act as boundaries to extract the translation error. We assume the existence of at least one common chunk between R and C. In cases where wrong translations can not be detected, the percentage of missing terms is reported as additional indicator[11].

3.4. Integrated CoTERM

We define here an integrated metric CoTERM, based on CoTERM$^+$ and CoTERM$^-$. A natural way is to take a linear combination of these two. The choice of coefficients is theoretically an important issue, and different possibilities exist. We adopt a theoretically radical viewpoint and set up a dividing condition based on the relation between the consistency of correct translations and the consistency of incorrect translations. If the former is larger, we reward the consistency of incorrect translations, while if the former is smaller, we penalize it. Based on this dichotomy, by denoting $a = \text{CoTERM}^+$ and $b = \text{CoTERM}^-$, the integrated CoTERM is represented as follows:

$$\text{CoTERM} = \begin{cases} \frac{1}{2}(a+b) & \text{if a > b} \\ \frac{1}{2}(a-b) & \text{if a} \leq \text{b} \end{cases} \quad (8)$$

CoTERM ranges between -100 to 100, where 100 means perfect consistency in term translation with no errors and -100 means consistent errors with no correct translations. Positive sign of CoTERM ($a > b$) means high consistency of correct translations and errors. Negative sign of CoTERM means the system is more consistent in making errors than producing correct translations. The sign of CoTERM is thus a strong indicator to diagnose, at the same time, system performance with regard to correctness and errors in term translation.

[11]This percentage is very low in our experiments.

4. Experimental Setups

We carried out a set of experiments for English-French, English-Japanese and Japanese-English. We focus on English-French and English-Japanese. Experiments on the Japanese-English direction are reported in the appendix A.3.

4.1. Data

For English-French, we used the Tico-19 test set used in WMT 2021 for terminology evaluation[12]. The Tico-19 benchmark includes 30 documents (3,071 sentences) translated from English into 36 languages[13]. The test set consists of 2,100 parallel sentences with terminology annotations. For English-Japanese, we used a test set provided in the Workshop on Asian Translation (WAT)[14]. The test set was from the ASPEC-JE Parallel corpus[15] (Nakazawa et al., 2016), which contains abstracts of 24 scientific domains. Unlike Tico-19, ASPEC does not have a consolidated document level and does not provide terminology annotations. We automatically annotated the occurrences of terms in the test set using a number of terminological resources in various fields including various science domains, technology (Kotani and Kori, 1990), information (Aiso, 1993) and economics (Nikkei, 2000), and then made them validated by a knowledgeable annotator. This led to a list of 1,300 sentences with term annotations.

The number of occurrences and the number of distinct terms are given in Table 2.

Test set	# Terms	# Distinct
Tico-19 (En)	2,557	215
Tico-19 (Fr)	5,879	324
ASPEC (En)	4,288	2,397
ASPEC (Ja)	5,656	2,899

Table 2: Statistics of Tico-19 and ASPEC tests.

4.2. NMTs and LLMs

For English-French we evaluate six NMT models including two transformers: transformer.wmt14.en-fr (Transf) and OPUS-MT (Opus) (Tiedemann and Thottingal, 2020), three convolutional models: conv.wmt14.en-fr (Conv) (Gehring et al., 2017), Lightconv.glu.wmt14.en-fr (LightC) and dynamicconv.glu.wmt14.en-fr (DynC) (Wu et al., 2019a) and the multilingual model NLLB (NLLB Team et al., 2022). We also evaluate seven LLMs: T5-Large (T5-L) (Raffel et al., 2020), MBart (Liu et al., 2020), M2M100 (Fan et al., 2021), Madlad (Kudugunta et al., 2023), Llama4 (Touvron et al., 2023), Gemma3 (Lieberum et al., 2024) and Qwen2 (Yang et al., 2024). For English-Japanese we use four transformers: Fugumt and Tako (Junczys-Dowmunt et al., 2018), Tatoeba (Tiedemann, 2020) and Wide-En-X (W-X) (Tran et al., 2021a). The used LLMs are the same except for T5-L which is replaced by Gemma2. For each LLM, we chose the best version based on BLEU.

4.3. Evaluation Metrics

As standard metrics, we use BLEU and character n-gram F-score (chrF) (Popović, 2015), as well as word embedding metrics: BertScore (Zhang et al., 2020) and COMET (Rei et al., 2020). For terminology-oriented evaluation, we use the Exact Match accuracy (EM), Window Overlap (WO2)[16] (ibn Alam et al., 2021a) and F1 and TC (Semenov et al., 2023). Finally, we use our proposed CoTERM metric at the document and corpus levels.

5. Results

Table 3 shows the results of NMT and LLM models on the Tico-19 (En-Fr) test set. Overall, in both generic and term-based metrics, LLMs models show better results than NMT models (Except NLLB and Transf). NLLB significantly[17] outperforms other models with BLEU score of 48.20. According to BLEU and BertScore, NLLB is ranked first, closely followed by Llama4. Other sentence-level metrics give different rankings. COMET ranked Gemma3 first and Madlad second while chrF ranked Llama4 first followed by Gemma3 with a thin margin. Regarding term-based measures, EM and WO2 agreed that Llama4 and NLLB are the two top MT systems but in a reverse order. The same observation can be made for F1 and TC but this time for Llama4 and Gemma3. We also observe differences in rankings between sentence-based and term-based metrics for LLMs and NMT that have little difference in BLEU. At the corpus level CoTERM$_f^+$ and CoTERM$_r^+$ agreed to rank Llama4 first and Madlad second. CoTERM$_f^+$ at the document level also ranked Llama4 first, though Gemma3 is given a second position. CoTERM$_r^+$ is basically in line with its flexible version though the order of top two is in reverse. It is also interesting to note differences in consistency between document-based and corpus-based values. At corpus level, T5-L shows better CoTERM$_r^+$ than MBart, while the contrary happens at the document level. If we

[12] https://tico-19.github.io/

[13] Japanese is not among these 36 languages.

[14] https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020

[15] https://jipsti.jst.go.jp/aspec

[16] We only report it for English-French data as Japanese is not handled.

[17] See Table 7 for significance test.

	En → Fr												
	NMT						LLM						
	Transf	Conv	LightC	DynC	Opus	NLLB 3B	T5-L 770M	MBart 610M	M2M 1B	Madlad 7B	Llama4 109B	Gemma3 27B	Qwen2 7B
BLEU	42.94	28.23	36.94	29.18	44.00	**48.20**	41.71	40.66	42.21	45.52	47.45	47.43	36.36
chrF	68.06	62.53	65.94	64.43	69.04	**72.22**	66.32	67.05	62.24	70.96	71.91	71.92	64.94
BertScore	90.79	90.66	90.68	90.84	91.13	**91.89**	90.69	90.56	90.67	91.23	91.77	91.60	89.44
COMET	86.06	85.22	85.66	85.95	85.98	87.28	85.88	85.45	85.93	87.56	87.46	**87.89**	83.60
EM	85.60	83.94	81.83	81.11	91.64	96.00	85.68	92.25	94.69	95.54	**96.31**	95.79	89.26
WO2	20.36	20.16	20.05	20.38	20.18	**25.31**	20.22	20.90	22.71	23.38	24.79	24.27	20.37
F1	72.70	74.26	71.99	71.45	76.26	87.01	70.62	76.65	84.20	81.54	87.28	**88.85**	79.97
TC	65.52	56.06	63.12	61.59	65.97	68.42	60.85	58.03	61.74	70.42	**72.55**	71.01	55.76
CoTERM$_f^+$	67.02	56.25	62.79	61.39	67.79	74.08	61.21	60.36	64.51	74.44	**75.93**	72.20	55.11
CoTERM$_r^+$	56.68	45.70	52.10	51.62	56.15	62.06	52.15	50.66	53.75	62.26	**64.04**	61.07	47.42
CoTERM$_f^+$d	77.04	66.97	73.18	71.93	80.32	83.80	74.85	77.61	81.50	84.12	**84.82**	84.48	73.77
CoTERM$_r^+$d	68.06	58.51	63.70	65.72	69.87	75.44	65.18	68.55	73.61	74.69	76.21	**76.79**	66.77
CoTERM$^-$	30.62	31.50	34.26	33.07	43.85	41.26	34.49	25.46	35.07	41.58	42.66	**46.45**	34.81
CoTERM$_f$	48.82	43.87	48.52	47.23	55.82	57.67	47.85	42.91	49.79	58.01	59.29	**59.32**	44.96
CoTERM$_r$	43.65	38.60	43.18	42.34	50.00	51.66	43.32	38.06	44.41	51.92	53.35	**53.76**	41.11
CoTERM$_f$d	53.83	49.23	53.72	52.50	62.08	62.53	54.67	51.53	58.28	62.85	63.74	**65.46**	54.29
CoTERM$_r$d	49.34	45.00	48.98	49.39	56.86	58.35	49.83	47.00	54.34	58.13	59.43	**61.62**	50.79

Table 3: Results of the evaluation of NMTs and LLMs on the Tico-19 test set for English-French. The best models are in bold, and the second bests underlined.

look at CoTERM$^-$, which measures the consistency of errors, we see that Gemma3 is the most consistent model in making errors while MBart is the most inconsistent. We also note that Opus is the second best consistent model for errors. The integrated CoTERM that takes into account both correct and incorrect consistencies show more agreement between corpus- and document-levels by ranking Gemma3 first and Llama4 second. Finally, if NLLB is the best system based on BLEU, it is not the best at handling terms as it is outperformed by Gemma3, Llama4 and Madlad as shown by CoTERM.

Table 4 reports the results of experiments on the ASPEC (En-Ja) test set. Overall, the results are much lower than En-Fr in all the metrics, which shows that translating from English to Japanese is a more difficult task than translating from English to French. We first note similar BLEU scores for Fugumt and W-X, with no significant difference. Except for BLEU and chrF, existing embedding and term-level measures ranked W-X first. There is, however, a disagreement in the second best system. BertScore, F1 and TC ranked Llama4 at the second position while COMET and EM ranked Gemma3. Our positive flexible and rigid CoTERM measures are in line with most of the measures by ranking W-X first. However, CoTERM$_f^+$ shows that Fugumt is the second best model. Regarding CoTERM$^-$, we see that M2M is the most consistent model, followed by Tako. W-X is ranked fourth. Finally, similarly to BLEU, both CoTERM$_f$ and CoTERM$_r$ ranked Fugumt first. Qwen2 is ranked second. Negative scores of CoTERM indicate that MT systems have higher consistency in errors while at the same time making more erroneous term translations than correct ones.

6. Human Evaluations

In order to observe the status of CoTERM in relation to human evaluations, we asked three native Japanese speakers, specialized in translation and language studies, to judge the subset of Japanese sentences generated by five systems randomly extracted from the ASPEC-En-Ja test set. Thus, 1,250 sentences have been annotated [18]. All the raters used the same website made for the evaluation, in which system translations are randomly ordered. All of them saw the sentence sets in the same order. Three scores are assigned to each sentence: 1) overall sentence score; 2) rigid term translation score and 3) flexible term translation score. Each sentence was rated from 1 to 5: 1 being very bad, i.e. all or most terms are incorrect; 2 bad, i.e. more than 50 percent of the terms are incorrect; 3 acceptable, i.e. around 50 to 70 percent of the terms are correct; 4 good, i.e. most terms are correct; and 5 being very good, i.e. perfect.

Table 5 shows Inter Annotators Agreement (IAA) in kappa (κ) (Cohen, 1960) and Pearson (Pearson, 1962) correlations for each rater pairs. The IAAs are high for all the pairs, indicating that the overall

[18] 250 source and reference sentences are paired with their five translations.

	En → Ja												
	NMT					LLM							
	Fugumt	Tako	Tatoeba	W-X	NLLB 3B	M2M 12B	mBART 610M	Gemma2 27B	Gemma3 27B	Llama3 70B	Llama4 109B	Qwen2 7B
BLEU	**27.53**	21.02	11.05	27.30	17.84	14.37	19.21	23.86	23.92	20.98	24.97	23.39
chrF	37.15	31.22	21.56	37.97	27.66	26.38	29.99	34.96	35.34	32.91	32.26	**38.24**
BertScore	86.25	84.57	77.87	**87.15**	82.33	78.86	83.97	85.81	86.21	85.53	86.48	85.98
COMET	88.80	86.30	80.11	**90.27**	85.60	83.11	86.90	89.82	90.22	89.18	89.77	86.99
EM	92.23	91.10	80.04	**93.83**	88.51	89.72	91.27	92.12	92.49	89.78	92.36	86.23
F1	55.00	51.18	38.50	**56.62**	47.62	46.81	50.49	55.11	55.63	52.71	55.76	50.80
TC	40.10	35.53	25.87	**41.24**	32.31	32.13	35.28	39.98	40.26	38.38	40.50	38.03
CoTERM$^{+}_{f}$	41.28	36.80	26.66	**42.17**	33.14	33.02	37.07	40.56	40.62	38.39	41.11	36.00
CoTERM$^{+}_{r}$	38.83	33.99	23.99	**39.78**	30.89	30.69	34.21	38.35	38.70	36.26	38.96	35.02
CoTERM^{-}	42.17	48.09	40.42	45.95	46.54	**50.01**	45.60	44.30	43.42	42.53	44.42	38.67
CoTERM$_{f}$	**-0.44**	-5.64	-6.88	-1.89	-6,70	-8.49	-4.26	-1.87	-1.40	-2.07	-1.65	-1.33
CoTERM$_{r}$	**-1,67**	-7.05	-8.21	-3.08	-7.82	-9.66	-5.69	-2.97	-2.36	-3.13	-2.73	-1.82

Table 4: NMTs and LLMs results for English-Japanese (Aspec). Best in bold, and second bests underlined.

	Kappa (κ)			**Pearson**		
	A12	A13	A23	A12	A13	A23
Term level						
Flexible	0.50	0.69	0.50	0.81	0.91	0.82
Rigid	0.54	0.73	0.52	0.82	0.93	0.83
Sent level	0.21	0.18	0.14	0.62	0.63	0.57

Table 5: Inter-Annotator Agreement. A12 means agreement between the raters A1 and A2.

Metric	**Kendall**			**Pearson**		
	A1	A2	A3	A1	A2	A3
CoTERM^{r}	0.43	0.41	0.43	0.49	0.46	0.48
CoTERM^{f}	0.47	0.45	0.46	0.57	0.51	0.55

Table 6: Kendall and Pearson correlations with human judgments on ASPEC En-Ja test set.

rating is reliable. The IAAs for term level rating is higher than the IAAs for the sentence level rating.

As CoTERM is calculated based on individual terms, the correlation between human raters and CoTERM need to be computed at the term level. In order to do so, for each set of sentences in which a term occurs, we compute CoTERM^{+} and the mean average score of the raters. The ASPEC sample contained 309 En-Ja term pairs. Table 6 shows the correlations in Kendall (Kendall and Smith, 1939) and Pearson correlation coefficients. We see moderate to strong correlations between human raters and CoTERM. Fleiss Kappa for all the raters are: CoTERM^{f} = 0.56 and CoTERM^{r} = 0.59. This indicates, though indirectly, that CoTERM reflects evaluations of term translations by humans.

7. Discussion

We introduced CoTERM as a set of consistency-oriented metrics of term translations based on HHI. It is an extension of C_t originally designed for evaluating human translations (Itagaki et al., 2007), fully extended to deal with MT systems in flexible and rigid scenarios at the segment level. We recognize similarities of CoTERM^{r} with TC (Semenov et al., 2023) in the sense that both compute consistency of terms. However, they differ in several points: i) TC score is based on the percentage of correct occurrences for each term, while CoTERM is based on HHI. ii) TC does not deal with multiple correct translations (instead it uses the first occurrence or the most frequent one). By using HHI, CoTERM allows the evaluation where multiple correct translations are allowed (flexible scenario). In addition, iii) CoTERM is designed to take into account the structure of the corpus at different levels: paragraph, document, corpus or any other unit of segment that needs to be defined.

Addressing the translation error consistency, i.e. CoTERM^{-}, is also the originality of CoTERM. As shown in the experiments, CoTERM^{-} tends to give different rankings from the other metrics and the best systems are not always the most consistent from the point of view of error translations. Finally, CoTERM can be easily interpreted based on its sign. If the score is positive (Table 3), this implies that consistency for correct translations is higher than errors while a negative sign (Table 4) indicates the contrary and reveals not only that the MT system is more consistent in making erroneous translations but also that the MT system is probably making more errors than correct term translations.

Extensive evaluations in En-Fr and En-Ja, given in Tables 3 and 4, show that best NMT and LLM systems in terms of BLEU, BertScore or COMET do not

necessarily treat terms consistently. This finding thus indicates the necessity of using fit-for-purpose measures that reflect requirements for specific aspects of translations.

Although our experiments have clearly shown that CoTERM metrics shed light on different aspects from generic MT metrics, it is important to explore further the relationships between accuracy and consistency reflected in CoTERM in both rigid and flexible setups. While CoTERM's merit resides in the fact that it gives a single metric that takes into account the basic requirements for handling terms in translation, the relationships between accuracy and consistency in CoTERM, especially within the flexible setup, is yet to be systematically analyzed.

Among the corpora used in this study, Tico-19 has document-level information, while ASPEC does not. We observed the correlations between CoTERM and human evaluations carried out at sentence level for ASPEC corpus. We show in the appendix some additional results of experiments, behaviour of CoTERM for some terms in Tico-19 and the correlation between human annotations and CoTERM. In order to fully explore the characteristics of CoTERM in different setups with regard to the level of documents and corpora, it is essential to carry out further human evaluations for individual terms with regard to their occurrences within a document or a corpora, taking into account the distributions of terms. This will enable us to further refine the overall CoTERM defined in equation (8).

An additional technical limitation should be noted here related to CoTERM^{-}, which was introduced for the first time as consistency evaluation for translation errors. In automatically calculating CoTERM^{-}, we adopted a heuristic algorithm under the assumption that reference and prediction sentences share sufficient amount of similar non-term sequences to locate term translation errors in the prediction sentence. Unfound term errors are marked as "unknown" and are excluded from the computation of CoTERM^{-}. This specific case needs to be addressed, although in reality it represents very few cases in our experiments.

8. Conclusion

In this paper, we introduced a new set of measures and a framework to evaluate MT systems with regard to their ability to properly translate terms. By using CoTERM together with existing standard and terminology-oriented measures, we empirically observed several pretrained NMT and LLM models with respect to their ability in translating terms, which, to the best of our knowledge, is the first work for a systematic comparison. One important finding lies in the fact that better scores in such generic metrics as BLEU or BertScore do not necessarily mean better performance in term translations. This indicates the importance of using fit-for-purpose evaluation metrics. We have shown that rigid and flexible CoTERM provide a means to compare MT models with respect to term translations. As future work, we will conduct evaluation on MT systems which have been trained to deal with terminology and will also explore the incorporation of CoTERM as part of a loss function for fine-tuning.

9. Limitations

In relation to the framework in which CoTERM metrics are applied, we assumed an annotated reference list where each term in the source sentence to translate is provided with its corresponding target term translations in the appropriate position. While Tico-19 and ASPEC-JE with automatically annotated terms could be used for our immediate purpose, it is important to define a detailed guideline for annotating term translations that reflects proper treatment of terms in translation and to construct corpora with appropriate term tags. In the process of human corrections of automatic term tagging to ASPEC-JE corpus, we defined a basic guideline for annotating term tags, but it remains tentative. In order to define such a guideline, it is essential to take into account existing work on term variations and how they are addressed in translation.

It is also required to establish an overall standard framework for evaluating term translations for MT systems. To facilitate work in this direction, we developed and made available a web-based interface for human raters to carry out generic and terminology-oriented evaluations. But to fully establish the framework is left for our further work.

Separately, in releasing our term annotation of the ASPEC dataset, there is a practical limitation i.e. we can only provide it to those who obtain the right to use ASPEC original datasets. Obtaining the right to use ASPEC original datasets is, however, straightforward.

10. Acknowledgments

This work is partly supported by JSPS S20110 and by JSPS 24H00736.

11. Bibliographical References

Kaori Abe. 2023. *Application-oriented Machine Translation: Design and Evaluation*. Ph.D. thesis, Tohoku University.

Hideo Aiso. 1993. *Terminological Dictionary of Information Processing*. Ohm, Tokyo.

Mateja Ambrožič, Mojca Jevšnik, and Peter Raspor. 2010. Inconsistent terminology in food safety field: A permanent risk factor? *Journal of Food and Nutrition Research*, 49(4):186–194.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jakob. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Anne Condamines. 2010. Variations in terminology: Applications to the management of risks related to language use in the workplace. *Terminology*, 16(1):30–50.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems.

Béatrice Daille, Benoît Habert, Christian Jacquemin, and Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Helmut Felber. 1984. *Terminology Manual*. UNESCO, Paris.

Sabela Fernández-Silva and Koen Kerremans. 2011. Terminological variation in source texts and translations: A pilot study. *Meta: Translator's Journal*, 56(2):318–335.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Judit Freixa. 2006. Causes of denominative variation in terminology: A typology proposal. *Terminology*, 12(1):51–77.

Judit Freixa. 2022. Causes of terminological variation. In Pamela Faber and Marie-Claude L'Homme, editors, *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, pages 399–420. John Benjamins.

Angelina Gašpar, Sanja Seljan, and Vlasta Kučiš. 2022. Measuring terminology consistency in translated corpora: Implementation of the Herfindahl-Hirschman index. *Information*, 13(2).

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2019. Termeval: An automatic metric for evaluating terminology translation in mt. In *CICLing 2019: The 20th International Conference on Computational Linguistics and Intelligent Text Processing*, Paris, France. European Language Resources Association (ELRA).

O.C. Herfindahl. 1950. *Concentration in the Steel Industry*. Columbia university.

Albert O. Hirschman. 1945. *National Power and the Structure of Foreign Trade*, 1 edition. University of California Press.

Md Mahfuz ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Md Mahfuz ibn Alam, Ivana Kavapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, and Kweon Woo Jung. 2021b. Findings of the wmt shared task on machine translation using terminologies. *Proc. of WMT 2021*, abs/2106.11891.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistenscy with statistical method. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kyo Kageura. 2021a. Automatic term processing in the context of translation: Theoretical and practical issues and prospects. In *8th Chinese Terminology Construction and Terminology and Cognition International Conference*.

Kyo Kageura. 2021b. The status of terms and concepts in the learned use of language: Invoking the wüsterian spirit in the era of machine learning. In Petra Drewer, Felix Mayer, and Donatella Pulitano, editors, *Terminologie: Industrie, Information, Intelligenz: Atken des Symposions*, pages 3–12. Deutsher Terminologie-Tag e.V.

Wioleta Karwacka. 2014. Quality assurance in medical translation. *The Journal of Specialised Translation*, 21(1):19–34.

M. G. Kendall and B. Babington Smith. 1939. The problem of m rankings. 10(3):275–287.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Froilan Gimenez Perez. 2024. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Online. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Takuya Kotani and Atsuhiko Kori. 1990. *Dictionary of Technical Terms*. Kenkyusha, Tokyo.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumatica*, 12(1):455–463.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth*

Conference on Machine Translation, pages 902–911, Singapore. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Nikkei. 2000. *Nikkei Dictionary of Economy and Business*. Nikkei, Tokyo.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

E. S. Pearson. 1962. Some thoughts on statistical inference. 33(2):394–403.

Katia Peruzzo. 2010. Horizontal denominative variation in an eu victim-related english-italian parallel corpus. *Rivista Internazionale di Tecnica della Traduzione*, 12(1):177–188.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192, Dublin, Ireland. European Association for Machine Translation.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fernando Prieto Ramos. 2020. Ensuring consistency and accuracy in legal terms in institutional translation: The role of terminological resources in international organizations. In Fernando Prieto Ramos, editor, *Institutional Translation and Interpreting: Assessing Practices and Managing for Quality*, pages 128–149. Routledge.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Stephen A. Rhoades. 1993. The Herfindahl-Hirschman index. *Federal Reserve Bulletin*, (Mar):188–189.

Margaret Rogers. 1997. Synonymy and equivalence in special-language texts: A case study in german and english texts on genetic engineering. In Anna Trosborg, editor, *Text Typology and Translation*, pages 217–245. John Benjamins.

Margaret Rogers. 2007. Terminological equivalence in technical translation: A problematic concept? st. jerome and technical translation. *Synaps*, 20(1):13–25.

Margaret Rogers. 2008. Terminological equivalence: Probability and consistency in technical translation. *LSP Translation Scenarios: MuTra Journal*, pages 101–108.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor. 2023. Findings of the wmt 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5877–5886. PMLR.

Gemma Team. 2024. Gemma.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021a. Facebook ai wmt21 news translation task submission. *arXiv:2108.03265v1 [cs.CL]*.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021b. Facebook ai's wmt21 news translation task submission. In *Proc. of WMT 2021*.

J.P. Woodard and J.T. Nelson. 1982. An information theoretic measure of speech recognition performance.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019a. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Lijun Wu, Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019b. Depth growing for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5558–5563, Florence, Italy. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A. MT Evaluation

A.1. NMT Models

Transformers We used the transformer model introduced in Ott et al. (2018). Two transformers models have already been tested on the Tico-19 test set that is: OPUS (Tiedemann and Thottingal, 2020) and FAIRSEQ (Ott et al., 2018). We refer to this model by Transf.

Convolutional Models The first convolutional model that we refer to as Conv, is the model proposed by Gehring et al. (2017). It is based entirely on convolutional neural networks and can be fully parallelized during training to better exploit GPU and optimization.

Lightweight and Dynamic Convolutions Wu et al. (2019a) introduced a lightweight and a dynamic convolutional networks. They are much simpler and more efficient than the self-attention convolutional models. We refer to these two models by LightC and DynC.

Opus-MT that we refer to as OPUS, is a transformer-based model (Tiedemann and Thottingal, 2020) trained on a large bitext corpora extracted from the web[19]. OPUS is used for English-French and Japanese-English.

Wide-X-En and Wide-En-X are two large multilingual models (Tran et al., 2021a) trained respectively to translate texts from multiple languages (X) to English (Wide-X-En[20]) and from English to multiple languages (Wide-En-X[21]).

Fugumt is a transformer model[22] used for both English to Japanese and Japanese to English directions. It uses the Marian-NMT library (Junczys-Dowmunt et al., 2018).

Tatoeba is an English to Japanese translation model. It is part of the Tatoeba challenge[23].

TakoMT is an X to Japanese translation model (X = de, en, es, fr, it, ru, uk) that uses the Marian-NMT library (Junczys-Dowmunt et al., 2018).

A.2. Evaluation Scores and Significance

Evaluating systems using BLEU is straightforward, as we can apply one of the existing evaluation scripts. However, it is important to pay attention to some technical details that may affect the evaluation. Some scripts require tokenized inputs, and different tokenization tools have a direct impact on the BLEU score which may compromise the process of comparing different NMT systems that use different tokenizers. For evaluation, we use SacreBleu[24] which alleviates the drawbacks of previous scripts such as multi-Bleu, etc. SacreBleu computes the BLEU score on detokenized inputs. Recent NMT systems provide detokenized outputs and in that case there is no detokenization problem.

[19] http://opus.nlpl.eu.

[20] https://huggingface.co/facebook/wmt21-dense-24-wide-x-en

[21] https://huggingface.co/facebook/wmt21-dense-24-wide-en-x

[22] https://github.com/s-taka/fugumt

[23] https://github.com/Helsinki-NLP/Tatoeba-Challenge

[24] https://github.com/mjpost/sacrebleu

	Tico-19 (En-Fr)	
System	**BLEU** ($\mu \pm$ 95% CI)	**chrF2** ($\mu \pm$ 95% CI)
Baseline NLLB (3B)	48.20 (48.20 $\pm$ 0.7945)	72.22 (72.2401 $\pm$ 0.4690)
Gemma3 (27b)	47.43 (47.43 $\pm$ 0.7849) (p = 0.0060)*	71.92 (71.9266 $\pm$ 0.4747) (p = 0.0290)*
Llama4	47.45 (47.4605 $\pm$ 0.8191) (p = 0.0030)*	71.91 (71.9274 $\pm$ 0.4720) (p = 0.0190)*
M2M100 (1.2B)	42.21 (42.2241 $\pm$ 0.7951) (p = 0.0010)*	68.24 (68.2569 $\pm$ 0.5656) (p = 0.0010)*
Madlad (7b)	45.52 (45.5312 $\pm$ 0.7904) (p = 0.0010)*	70.96 (70.9768 $\pm$ 0.4786) (p = 0.0010)*
mBART	40.66 (40.6740 $\pm$ 0.8181) (p = 0.0010)*	67.05 (67.0685 $\pm$ 0.5958) (p = 0.0010)*
Qwen2 (7B)	36.36 (36.3626 $\pm$ 0.7257) (p = 0.0010)*	64.94 (64.9497 $\pm$ 0.4577) (p = 0.0010)*
T5-large	41.71 (41.7299 $\pm$ 0.8618) (p = 0.0010)*	66.75 (66.7747 $\pm$ 0.7227) (p = 0.0010)*
Opus	44.00 (44.0134 $\pm$ 0.8323) (p = 0.0010)*	69.04 (69.0578 $\pm$ 0.6842) (p = 0.0010)*
Conv	28.23 (28.2437 $\pm$ 0.6665) (p = 0.0010)*	62.53 (62.5427 $\pm$ 0.5932) (p = 0.0010)*
DynConv	29.18 (29.1947 $\pm$ 0.7022) (p = 0.0010)*	64.43 (64.4425 $\pm$ 0.5569) (p = 0.0010)*
LightConv	36.94 (36.9487 $\pm$ 0.7725) (p = 0.0010)*	65.94 (65.9576 $\pm$ 0.5738) (p = 0.0010)*
Transformer	42.94 (42.(9309 $\pm$ 0.7929) (p = 0.0010)*	68.06 (68.0816 $\pm$ 0.5533) (p = 0.0010)*

Table 7: Test significance over En-Fr MT systems based on bootstrap resampling over 1000 samples.

	ASPEC (En-Ja)	
System	**BLEU** ($\mu \pm$ 95% CI)	**chrF2** ($\mu \pm$ 95% CI)
Baseline Fugumt	27.5319 (27.5314 $\pm$ 0.9744)	37.1537 (37.1515 $\pm$ 0.9145)
Gemma2 (27B)	23.8636 (23.8668 $\pm$ 0.8162) (p = 0.0010)*	34.9653 (34.9678 $\pm$ 0.8258) (p = 0.0010)*
Gemma3 (27B)	23.9252 (23.9264 $\pm$ 0.8228) (p = 0.0010)*	35.3420 (35.3423 $\pm$ 0.7662) (p = 0.0010)*
llama3.1 (70B)	21.0612 (21.0589 $\pm$ 0.7550) (p = 0.0010)*	32.7147 (32.7105 $\pm$ 0.7531) (p = 0.0010)*
Llama4	24.9751 (24.9771 $\pm$ 0.8773) (p = 0.0010)*	36.2654 (36.2735 $\pm$ 0.8186) (p = 0.0070)*
M2M100 (12B)	14.3741 (14.3664 $\pm$ 0.6178) (p = 0.0010)*	26.3811 (26.3690 $\pm$ 0.7360) (p = 0.0010)*
mbart O2M	19.1951 (19.1822 $\pm$ 0.8011) (p = 0.0010)*	30.5172 (30.5116 $\pm$ 0.8172) (p = 0.0010)*
NLLB (3B)	17.8496 (17.8555 $\pm$ 0.7914) (p = 0.0010)*	27.6655 (27.6675 $\pm$ 0.7350) (p = 0.0010)*
Qwen2 (7B)	23.3974 (23.3857 $\pm$ 0.9941) (p = 0.0010)*	38.2484 (38.2338 $\pm$ 1.0179) (p = 0.0380)*
Tako	21.0279 (21.0290 $\pm$ 0.8746) (p = 0.0010)*	31.2227 (31.2243 $\pm$ 0.8405) (p = 0.0010)*
Tatoeba	11.0556 (11.0633 $\pm$ 0.6362) (p = 0.0010)*	21.5653 (21.5741 $\pm$ 0.6293) (p = 0.0010)*
W-X	27.3008 (27.3084 $\pm$ 0.8988) (p = 0.1958)	37.9791 (37.9864 $\pm$ 0.8880) (p = 0.0100)*

Table 8: Test significance over En-Ja MT systems based on bootstrap resampling over 1000 samples.

	ASPEC (En-Ja)
System	**CoTERM** ($\mu \pm$ 95% CI)
Fugumt (Baseline)	41.28
Gemma2	40.56 (p = 0.0010)*
Tako	36.80 (p = 0.0010)*
Tatoeba	26.66 (p = 0.0015)*
W-X	42.17 (p = 0.7192)

Table 9: Significance Test for CoTERM (En-Ja) based on bootstrap resampling over 1000 samples.

For systems that provide tokenized outputs, we apply the same tokenizer provided by mosesdecoder. We also use BertScore (Zhang et al., 2020) and COMET (Rei et al., 2020). Tables 8 and 9 report the test significance over the five tested models on the ASPEC En-Ja test set using bootstrap resampling method over 1000 samples. We see comparable significance results between BLEU and CoTERM where Fugumt significantly outperforms Gemma2, Tako and Tatoeba while the difference in score with W-X is not significant.

A.3. Additional Results

We report additional results for each model (Tables 10 to 16). The overall results for Japanese to English are reported in Table 17.

	T5 (En → Fr)				
	Small	Base	Large	3B	11B
BLEU	37.98	41.18	**41.71**	40.85	40.37
chrF	64.59	66.32	**66.75**	66.30	65.86

Table 10: T5 results on Tico-19.

	NLLB			
	0.6M	1B	1B-dist	3B
BLEU (En-Fr)	44.14	46.54	47.04	**48.20**
chrF (En-Fr)	69.83	71.17	71.61	**72.22**
BLEU (En-Ja)	11.27	14.58	13.85	**17.84**
chrF (En-Ja)	24.50	24.74	26.17	**27.66**
BLEU (Ja-En)	9.26	9.05	**13.20**	10.47
chrF (Ja-En)	37.33	34.88	**45.11**	39.00

Table 11: NLLB results on Tico-19 and ASPEC.

	Madlad (En → Fr)		
	3B	7B	10B
BLEU	42.04	**45.52**	41.67
chrF	70.83	**70.96**	70.55

Table 12: Madlad results on Tico-19.

	Llama				
	V3.1		V3.2	V3.3	V4
	7B	70B	3B	70B	109B
BLEU (En-Fr)	37.98	44.53	32.37	46.50	**47.45**
chrF (En-Fr)	66.00	70.35	62.10	71.33	**71.91**
BLEU (En-Ja)	14.27	21.06	8.49	20.98	**24.97**
chrF (En-Ja)	25.33	32.71	16.27	32.91	**36.26**
BLEU (Ja-En)	12.71	16.93	9.73	**18.77**	17.90
chrF (Ja-En)	47.90	52.88	44.42	**54.63**	54.20

Table 13: Llama results on Tico-19 and ASPEC.

	Gemma					
	7B	9B	27b	4B	12B	27B
BLEU (En-Fr)	34.18	44.41	45.70	43.36	46.18	**47.43**
chrF (En-Fr)	62.28	70.33	70.80	69.48	71.24	**71.92**
BLEU (En-Ja)	13.09	21.52	23.86	16.31	20.93	**23.92**
chrF (En-Ja)	23.71	32.39	34.96	28.47	32.86	**35.34**
BLEU (Ja-En)	11.42	16.56	16.88	14.54	15.89	**16.89**
chrF (Ja-En)	46.02	52.68	**53.41**	50.74	52.37	53.17

Table 14: Gemma results on Tico-19 and ASPEC.

	M2M100		
	418M	1B	12B
BLEU (En-Fr)	37.07	**42.21**	38.98
chrF (En-Fr)	65.11	68.24	**68.54**
BLEU (En-Ja)	11.08	13.82	**14.37**
chrF (En-Ja)	22.70	25.62	**26.38**
BLEU (Ja-En)	13.19	**14.71**	14.12
chrF (Ja-En)	45.20	**48.28**	47.40

Table 15: M2M100 results on Tico-19 and ASPEC.

	mBART				
	En → Fr		En → Ja		Ja → En
	O2M	M2M	O2M	M2M	M2M
BLEU	**40.66**	39.76	19.19	**19.21**	15.77
chrF	**67.05**	66.44	**30.51**	29.99	49.67

Table 16: mBART results on Tico-19 and ASPEC. O2M: means one to many and M2M: many to many.

B. Human Evaluation

B.1. Rating Guideline

For the Flexible and Rigid scores: rate the term translation quality of a given translation between 1 to 5. The term translation should be the same as the one used in the reference translation. Table 18 gives the definitions of the scores. For the Overall score: Rate the overall translation quality of a given translation between 1 to 5. A human translation is given for your reference. Table 19 gives the

	Ja → En										
	NMT			LLM							
	Fugumt	Opus	W-X	NLLB 3B	M2M 12B	mBART 610M	Gemma2 27B	Gemma3 27B	Llama3.3 70B	Llama4 109B	Qwen2
BLEU	**21.19**	9.99	16.40	10.47	14.71	15.77	16.88	16.89	18.77	17.90	13.02
chrF	54.17	39.13	53.08	39.00	48.28	49.67	53.41	53.17	**54.63**	54.20	50.69
chrF++	50.43	36.08	49.54	36.05	44.67	46.04	49.02	48.88	**50.54**	50.08	46.17
BertScore	93.18	90.37	92.59	89.53	92.22	92.77	93.36	93.35	**93.49**	93.32	91.70
COMET	77.33	68.13	75.00	71.53	79.21	81.18	**83.62**	83.38	83.54	83.16	80.03
EM	92.03	80.06	92.54	79.57	88.81	91.11	91.57	92.44	93.10	**93.23**	91.38
F1	69.84	44.51	71.05	47.31	70.80	62.45	70.80	69.72	71.92	**72.03**	68.05
TC	62.91	35.59	64.12	38.83	62.98	52.75	62.98	62.95	64.43	**64.70**	59.95
CoTERM$_r^+$	65.91	37.59	67.46	39.77	51.47	56.02	65.77	65.70	67.43	**67.89**	62.21
CoTERM$_r^+$	62.87	32.80	64.09	32.52	47.07	52.74	62.14	61.36	64.32	**64.86**	58.74
CoTERM$_f^-$	60.11	55.19	61.22	55.70	61.02	62.73	62.96	60.14	**63.46**	62.47	62.54
CoTERM$_f$	63.01	-8.80	64.34	-7.96	-4.27	-3.35	64.36	62.92	**65.44**	65.18	-0.16
CoTERM$_r$	61.49	-11.19	62.65	-11.59	-6.97	-4.99	-0.41	60.75	**63.89**	63.66	-1.90

Table 17: Results of the evaluation of NMTs and LLMs on the Aspec test set for English-French. The best models are in bold, and the second bests underlined.

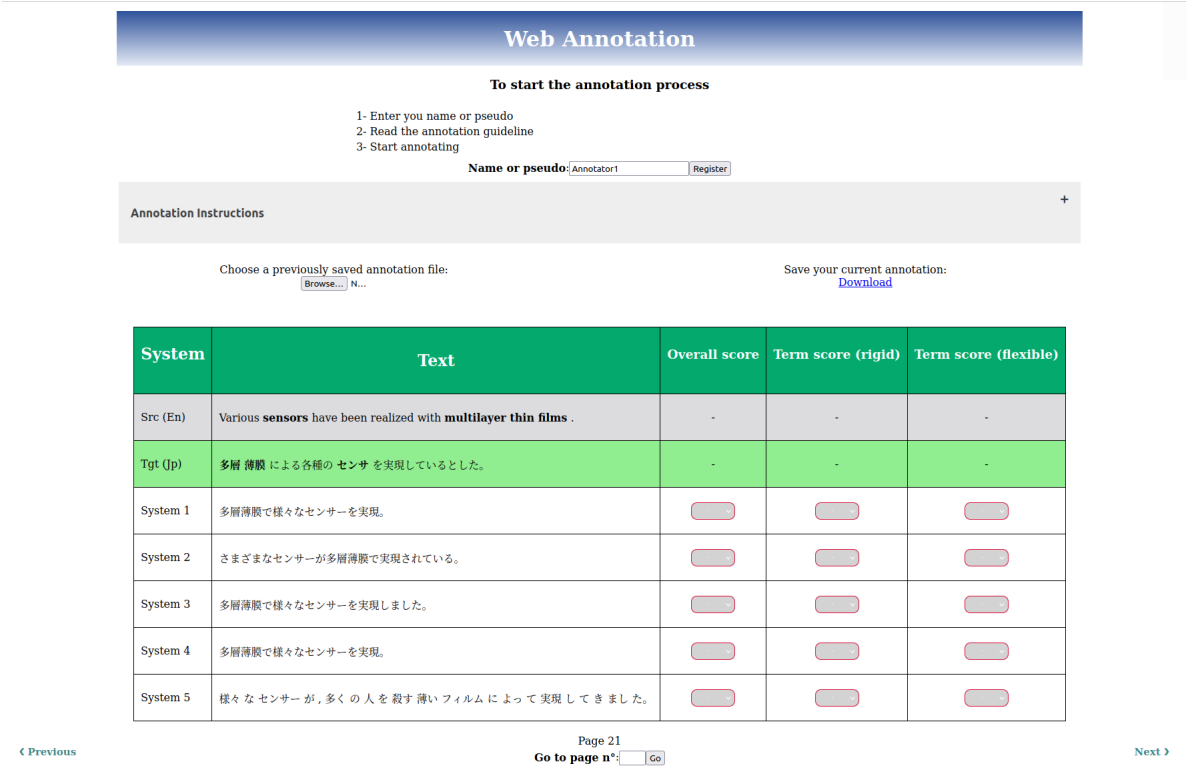

Figure 1: Web rating interface

definitions of the scores.

B.2. Inter-Annotator Agreement

We report in this section several experiments on pairwise raters agreement. Tables 20 and 21 report Kappa and Kendall scores respectively, for all the models (noted 5 systems) and for each model separately. In addition, Table 22 includes Mathew's Correlation Coefficient (MCC), Pearson and Spearman scores. We observe moderate to high correlation between annotators for term-based measures.

We report in Table 23 the Kendall (Kendall and Smith, 1939) score for MT systems on Tico-19. BLEU shows a significant higher correlation with 1-TER, chrF++ and BertScore than with CoTERM which suggests that consistency-based measures may bring a different perspective to MT evaluation.

Score	Label	Definition
1	very bad	all or most terms are incorrect
2	bad	more than 50 percent of the terms are incorrect
3	acceptable	around 50 to 70 percent of the terms are correct
4	good	most terms are correct
5	very good	perfect match

Table 18: Rating instructions for rigid and flexible translation scores.

Score	Label	Definition
1	very bad	incomprehensible, or comprehensible but the meaning is totally different from the original
2	bad	meaning is different from the original or misleading in critical points
3	acceptable	the core part of the original is comprehensible though some parts are not precise or missing
4	good	reasonably comprehensible but can be improved further
5	very good	perfectly comprehensible and can be used as is

Table 19: Rating instructions for overall translation score.

		Raters		
Systems	**Level**	A1-A2	A1-A3	A2-A3
5 systems	Term level			
	Flexible	0.50	0.69	0.50
	Rigid	0.54	0.73	0.52
	Sentence level	0.21	0.18	0.14
FuguMT	Term level			
	Flexible	0.54	0.73	0.53
	Rigid	0.57	0.77	0.56
	Sentence level	0.09	0.23	0.11
Tatoeba	Term level			
	Flexible	0.48	0.63	0.49
	Rigid	0.49	0.74	0.51
	Sentence level	0.30	0.24	0.16
Tako	Term level			
	Flexible	0.49	0.66	0.46
	Rigid	0.52	0.70	0.47
	Sentence level	0.13	0.04	0.07
W-En-X	Term level			
	Flexible	0.50	0.73	0.48
	Rigid	0.59	0.73	0.5
	Sentence level	0.09	0.14	0.04
Gemma2	Term level			
	Flexible	0.38	0.63	0.42
	Rigid	0.43	0.64	0.48
	Sentence level	0.17	0.06	0.12

Table 20: Inter-Annotator Agreement (kappa).

		Raters		
System	**Level**	A1-A2	A1-A3	A2-A3
5 systems	Term level			
	Flexible	0.74	0.86	0.74
	Rigid	0.75	0.88	0.74
	Sent level	0.55	0.57	0.50
FuguMT	Term level			
	Flexible	0.73	0.87	0.76
	Rigid	0.73	0.89	0.75
	Sent level	0.36	0.47	0.45
Tatoeba	Term level			
	Flexible	0.68	0.82	0.70
	Rigid	0.72	0.87	0.70
	Sent level	0.62	0.47	0.49
Tako	Term level			
	Flexible	0.71	0.85	0.70
	Rigid	0.69	0.84	0.68
	Sent level	0.41	0.40	0.36
W-En-X	Term level			
	Flexible	0.71	0.88	0.72
	Rigid	0.76	0.89	0.73
	Sent level	0.43	0.49	0.37
Gemma2	Term level			
	Flexible	0.65	0.79	0.64
	Rigid	0.64	0.85	0.67
	Sent level	0.52	0.60	0.37

Table 21: Inter-Annotator Agreement (kendall).

C. Examples

This appendix presents in section C.1 some translation examples produced by the five NMT models (Transf, Conv, LightC, DynC and OPUS) on the English-French Tico-19 test set. It also shows some translation examples of the NMT models tested on the English-Japanese ASPEC test set (Fugumt, Tako, Tatoeba and Wide-En-X) as well as the Japanese-English ASPEC test set (Fugumt, OPUS and Wide-X-En). Section C.2 presents some examples of CoTERM score calculations produced by Transf and OPUS models on several terms.

	Kappa			Fleiss			MCC			Pearson			Spearman		
Evaluation	A12	A13	A23	A12	A13	A23	A12	A13	A23	A12	A13	A23	A12	A13	A23
Term level															
Flexible	0.50	0.69	0.50	0.49	0.68	0.50	0.51	0.71	0.50	0.81	0.91	0.82	0.81	0.90	0.82
Rigid	0.54	0.73	0.52	0.53	0.72	0.52	0.56	0.74	0.53	0.82	0.93	0.83	0.82	0.92	0.82
Sent level	0.21	0.18	0.14	0.19	0.14	0.12	0.22	0.20	0.16	0.62	0.63	0.57	0.63	0.63	0.58

Table 22: Inter-Annotator Agreement (Cohen's kappa score (Kappa), Fleiss kappa (Fleiss), Mathews Correlation Coefficient (MCC), Pearson and Spearman.

	BLEU	CoTERM^f	CoTERM^r
BLEU	1.00	0.61	0.66
chrF	0.66	0.61	0.66
chrF++	0.87	0.64	0.59
BertScore	0.87	0.64	0.64
COMET	0.55	0.38	0.55
EM	-0.19	0.02	0.02
WO2	0.22	0.05	0.11
1-TER	0.94	0.66	0.72
CoTERM^f	0.61	1.00	0.61
CoTERM^r	0.66	0.61	1.00

Table 23: Kendall correlation score ($p < 0.05$ except for COMET, EM and WO2).

Level	**Fleiss Kappa**
Term level	
Flexible	0.56
Rigid	0.59
Sentence level	0.16

Table 24: Inter-Annotator Agreement for three raters (Fleiss kappa).

Measure	Interpretation
Cohen's kappa	
0	No agreement
0.10 - 0.20	Fair agreement
0.21 - 0.40	Slight agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement
Spearman & Kendall	
0.01 - 0.19	No or negligible relationship
0.20 - 0.29	Weak relationship
0.30 - 0.39	Moderate relationship
0.40 - 0.69	Strong relationship
$\geq$ 0.70	Very Strong relationship
Pearson & MCC	
0.0 - 0.19	Very low correlation
0.20 - 0.39	Low correlation
0.40 - 0.59	Moderate correlation
0.60 - 0.79	High correlation
0.80 - 1	Very high correlation

Table 25: Interpretation of correlation scores.

C.1. NMT System Examples

The following Tables 27, 28 and 29 illustrate some examples of predicted translations made by the used pre-trained models in our experiments. We emphasise correct term translations in green, variants or acceptable translations in blue and wrong translations in red color. The terms in the source reference (Src) and its target translation (Tgt) are marked in bold.

In the first example of Table 27, four out of five models failed in translating the term SARS-CoV-2. While in the second example, they all succeeded in translating *pandemic* and *SRAS*. That is said, the translation of other words, if semantically correct sometimes (*disparaître* and *mourra* for instance), leaves a room for improvement. In the last example of Table 27, all the NMT systems failed in translating the term: *ongoing outbreak* while the correct translation is *épidémie en cours*, some models correctly translate outbreak into *épidémie* but failed in finding the second part on the compound term that is *en cours*. However, the Conv and OPUS models for instance, translated *ongoing* by *actuelle* which is semantically correct. Hence, using the flexible CoTERM, the translation *épidémie actuelle* could be counted as correct, as long as *épidémie actuelle* is present in the reference list of correct terms.

As for Table 28, the two terms in first example are both properly translated in Fugumt and Wide-En-X. Tatoeba failed in both terms but its sentence structure is essentially the same as those in Fugumt and Wide-En-X. The second and third examples indicate that complex terms are translated into expressions that partially match the correct terms. The two terms in the last example are translated correctly in three systems, while Tatoeba misses one term completely. Some of the variants should be evaluated as correct by flexible CoTERM (CoTERM^f), although this ultimately depends on how we define the reference list because they cannot be ad-

Src Term (t)	Tgt Term	|t|	k	f_i	$(\frac{f_i}{k} \times 100)^2$	$\frac{C_t}{100}$	$\frac{\text{CoTERM}_t}{100}$
spread	propagation	3	2	1	2,500	50	22,22
	diffusion			1	2,500		
covid-19	covid-19	2	2	2	10,000	100	100
WHO	OMS	2	2	2	10,000	100	100
clean technologies	technologies propres	1	1	1	10,000	100	100
ongoing outbreak	épidémie en cours	1	0	0	0	0	0
pandemic	pandémie	1	1	1	10,000	100	100

Table 26: Illustration of Consistency Index (C_t) and CoTERM$_t$ scores calculation of Example 1.

.

dressed by lemmatizations. Table 29 shows some syntagmatic variants such as singular/plural and noun/adjective produced in different systems. The last example shows lexical variants. Examples of lexical variants in these Tables indicate the importance of designing proper evaluation dataset in fully taking advantage of CoTERM^f. This, we state in Limitation section, is one of our future tasks.

C.2. Examples of CoTERM Calculation

Table 26 shows the details of scores calculation of Example 1 (Section (3.3.1). Table 30 illustrates some terms and their correct translations generated by Transf and OPUS models. Both Tables report the occurrence of terms in the reference list (|t|), the total number of correct translations found by the model (k), the total number of correct translations found by the model for a given term (f), and finally, C_t score and CoTERM.

In the first example given by *SARS-CoV-2* in Table 30, we see that CoTERM score is quite high (6800 or 68%) even if the correct translations were found only 5 times out of 178 (4 times for *SRAS-CoV-2* and 1 time for *SARS-CoV-2*). By applying CoTERM, the score drastically drop (CoTERM$^{\text{w}}_{\text{f}}$ = 0.15 for instance) which reflects more the inconsistency of the Transf model with regard to the translation of *SARS-CoV-2*. On the contrary, when the model succeeds in most cases to translate a term, as we can see for the term *symptoms* for instance, the drop for the CoTERM scores is very low (we go from the perfect match of CoTERM = 10,000 to a little drop for CoTERM$^{\text{w}}_{\text{f}}$=9434.1). In this case, the drop scores reflect the two missed translations of the term *symptoms*. This statement can be observed in all the examples given in Table 30.

D. Special Remarks about LLMs

We tried to cover in our experiments most of the recent LLMs. We briefly discuss hereafter additional LLMs that have not been tested or that have been tested but not reported in our experiments. We have tested Deepseek-r1, however the translations were surprisingly very weak for the three tested directions (En-Fr, En-Ja and Ja-En). The translations often contained mixed languages, or were incomplete. Even by handling the <think> tag, which comes before every translation candidate, the obtained BLEU scores for several models were too low (for Ja-En direction for instance, we obtained: BLEU: 0.85, chrF: 13.31 and chrF++: 12.36 for Deepseek-r1:1.5b). In addition, the inference time was slow compared to other models. We let further explorations of DeepSeek-r1 and DeepSeek-v3 for future work. Qwen3 displays better translation outputs, however the inference time was too slow to be used for translating thousands of sentences. Similarly to Deepseek, Qwen3 uses the <think> tag in each output. We did not consider non-open source or free to use models such as chatGPT. If it is interesting to observe the behavior of leading models, recent work by (Moslem et al., 2023) have shown that NLLB outperformed chatGPT in MT. In this work, we have extensively tested NLLB. We believe that not testing chatGPT does not affect the takeaway message of our work.

	Tico-19 (En-Fr)
Src	The search for the animal origins of **SARS-CoV-2** is still on.
Tgt	les recherches sur les origines animales du **SARS-CoV-2** sont en cours.
Transf	La recherche de l'origine animale du CoV-2 du SRAS est toujours en cours.
Conv	La recherche des origines animales du CoV-2 du SRAS est toujours en cours.
LightC	La recherche sur les origines animales du CoV-2 du SRAS est toujours en cours.
DynC	La recherche des origines animales du CoV-2 du SRAS est toujours en cours.
OPUS	La recherche des origines animales du **SRAS-CoV-2** est toujours en cours.
Src	Will it become a **pandemic** worldwide, die down like **SARS** or relapse periodically like the **flu**?
Tgt	Va-t-il évoluer en **pandémie** mondiale, disparaître comme le **SARS** ou refaire surface régulièrement comme la **grippe** ?
Transf	Deviendra-t-elle une **pandémie** mondiale, disparaîtra-t-elle comme le **SRAS** ou rechutera-t-elle périodiquement comme la **grippe** ?
Conv	Deviendra-t-elle une **pandémie** dans le monde entier, mourra-t-elle comme le **SRAS** ou la rechute périodiquement comme la **grippe** ?
LightC	Deviendra-t-elle une **pandémie** mondiale, disparaîtra-t-elle comme le **SRAS** ou des rechutes périodiques comme la **grippe** ?
DynC	Deviendra-t-elle une **pandémie** mondiale, s'éteindra comme le **SRAS** ou rechutera périodiquement comme la **grippe** ?
OPUS	Se transformera-t-elle en **pandémie** mondiale, mourra-t-elle comme le **SRAS** ou rechutera-t-elle périodiquement comme la **grippe** ?
Src	Herein lie the secrets of why **asymptomatic** carriers are seen and what causes the severe cases in human **infection**.
Tgt	c'est ici que résident les secrets de l'observation de porteurs **asymptomatiques** et de ce qui cause des cas d'**infection** sévère chez l'homme.
Transf	C'est là que se trouvent les secrets de la présence de porteurs **asymptomatiques** et de la cause des cas graves d'**infection** humaine.
Conv	Voici les secrets des raisons pour lesquelles les porteurs **asymptomatiques** sont vus et ce qui cause les cas graves d'**infection** humaine.
LightC	C'est là que se trouvent les secrets de la détection des porteurs **asymptomatiques** et des causes des cas graves d'**infection** humaine.
DynC	On y trouve les secrets des raisons pour lesquelles on voit des porteurs **asymptomatiques** et les causes des cas graves d'**infection** chez les humains.
OPUS	Voici les secrets de pourquoi les porteurs **asymptomatiques** sont vus et ce qui cause les cas graves dans l'**infection** humaine.
Src	On Wednesday, the World Health Organization (**WHO**) declared the **ongoing outbreak** of **COVID-19** — the disease caused by **coronavirus SARS-CoV-2** — to be a **pandemic**.
Tgt	mercredi, l'Organisation mondiale de la santé (**OMS**) a qualifié l'**épidémie en cours** de **COVID-19** (la maladie provoquée par le **coronavirus SARS-CoV-2**) de **pandémie**.
Transf	Mercredi, l'Organisation mondiale de la Santé (**OMS**) a déclaré que l'**éclosion actuelle** de **COVID-19** de la maladie causée par le **coronavirus** CoV-2 SRAS-était une **pandémie**.
Conv	Mercredi, l'Organisation mondiale de la santé (**OMS**) a déclaré que l'**épidémie actuelle** de **COVID-19** est la cause de la **pandémie** de la maladie causée par le **coronavirus CoV-2-CoV-2**.
LightC	Mercredi, l'Organisation mondiale de la Santé (**OMS**) a déclaré que la **flambée** de **COVID-19**, causée par le **coronavirus CoV-2 du SRAS**, était une **pandémie**.
DynC	Mercredi, l'Organisation mondiale de la santé (**OMS**) a déclaré que l'**éclosion** en cours de **COVID-19**, la maladie causée par le **coronavirus CoV-2-SRAS**, est une **pandémie**.
OPUS	Mercredi, l ' Organisation mondiale de la santé (**OMS**) a déclaré que l ' **épidémie actuelle** de COVID-19 – la maladie causée par le **coronavirus SRAS-CoV-2** – était une **pandémie**.

Table 27: Translation examples obtained by the five tested NMT models (Transf, Conv, LightC, DynC and OPUS) on the English-French Tico-19 test set. Src and Tgt respectively correspond to the source sentence and Tgt its translation reference.

ASPEC (En-Ja)	
Src	Responding to these changes **DERS** can compute new **dose rate**.
Tgt	**DERS**はこれらの変化に対応して新たな線量率を計算できる。
Fugumt	これらの変化に対応して、**DERS**は新しい線量率を計算することができる。
Tako	これらの変化に対応する**DERS**は新しい用量率を計算することができます。
Tatoeba	**DAS** はこれらの変更に対応して,新しい量量を計算することができる.
Wide-En-X	これらの変化に対応して、**DERS**は新しい線量率を計算することができます。
Src	The **geometric-optical theory** of **standing wave** based on **ray** coincidence is presented.
Tgt	光線一致に基づく定常波の幾何光学的理論を展開した。
Fugumt	光線一致に基づく定在波の幾何光学ϕ理論を示す。
Tako	ビーム偶然に基づく立波の幾何学的・光学的理論を提案する。
Tatoeba	線の偶然による波の立っていることに関する幾何学的な理論が示されています。
Wide-En-X	線の一致に基づく定常波の幾何光学理論が提示される。
Src	The titled **measurement technology** on **vibration** and **temperature** in the heavy **machinery industry** is explained.
Tgt	重機械工業における振動と温度に関する標記計測技術を解説した。
Fugumt	重機産業における振動・温度測定技術について解説します。
Tako	重機産業における振動・温度測定技術について解説しています。
Tatoeba	重い機械産業における震動と温度に関する表向きの測定技術については説明がなされています。
Wide-En-X	と題し、重機業界における振動・温度の計測技術について解説する。
Src	The patient died of **respiratory insufficiency** caused by **pneumonia** 2 years later.
Tgt	約2年の経過後,肺炎による呼吸不全にて死亡した。
Fugumt	2年後に肺炎による呼吸不全で死亡した。
Tako	2年後に肺炎による呼吸不全で死亡。
Tatoeba	その患者は2年後に肺炎で死亡しました。
Wide-En-X	患者は2年後に肺炎による呼吸不全で死亡した。

Table 28: Translation examples obtained by the four tested NMT models (Fugumt, Tako, Tatoeba and Wide-En-X) on the English-Japanese ASPEC test set. Src and Tgt respectively correspond to the source sentence and Tgt its translation reference. ϕ indicates missing part.

.

	ASPEC (Ja-En)
Src	**DERS**はこれらの変化に対応して新たな線量率を計算できる。
Tgt	Responding to these changes **DERS** can compute new **dose rate**.
Fugumt	**DERS** can calculate new **dose rates** in response to these changes.
OPUS	The **DEARS** can calculate a new **linear rate** in response to these changes.
Wide-X-En	**DERS** can calculate a new **dose rate** in response to these changes.
Src	光線一致に基づく定常波の幾何光学的理論を展開した。
Tgt	The **geometric-optical theory** of **standing wave** based on **ray** coincidence is presented.
Fugumt	We developed a **geometrical optical theory** of **stationary waves** based on **ray**-matching.
OPUS	I've developed a **geometric theory** of **regular waves** based on a **beam of light**.
Wide-X-En	He developed a geometrically circular **geometrical optical theory** of **standing waves** based on the correspondence of light **rays**.
Src	重機械工業における振動と温度に関する標記計測技術を解説した。
Tgt	The titled **measurement technology** on **vibration** and **temperature** in the heavy **machinery industry** is explained.
Fugumt	This paper describes the characteristic **measurement technology** of **vibration** and **temperature** in heavy **machinery industry**.
OPUS	I've discussed **Marking techniques** for **vibrations** and **temperature** in heavy **machinery industries**.
Wide-X-En	The marking and **measuring technology** for **vibration** and **temperature** in the heavy **machinery industry** was explained.
Src	約2年の経過後,肺炎による呼吸不全にて死亡した。
Tgt	The patient died of **respiratory insufficiency** caused by **pneumonia** 2 years later.
Fugumt	Two years later, he died of **respiratory failure** due to **pneumonia**.
OPUS	About two years later, he died of **respiratory failure** due to **pneumonia**.
Wide-X-En	About two years later, he died of **respiratory failure** due to **pneumonia**.

Table 29: Translation examples obtained by the three tested NMT models (Fugumt, OPUS and Wide-X-En) on the English-Japanese ASPEC test set. Src and Tgt respectively correspond to the source sentence and Tgt its translation reference.

.

Transf model (En-Fr)									
Src Term	**Tgt Term**	**|t|**	**k**	**f**	$(\frac{f}{k} \times 100)^2$	C_t	**CoTERM**$_{\text{f}}$	C_t^w	**CoTERM**$_{\text{f}}^{\text{w}}$
SARS-CoV-2	SRAS-CoV-2	178	5	4	6400.0	6800.0	5.37	191.01	0.15
	SARS-CoV-2			1	400.0				
virus	virus	164	162	159	9633.06	9636.49	9402.89	9518.97	9288.22
	viral			3	3.43				
	virale			0	0.0				
	virales			0	0.0				
symptoms	symptômes	104	102	102	10000.0	10000.0	9619.08	9807.69	9434.1
outbreak	épidémie	88	30	29	9344.44	9355.55	1087.29	3189.39	370.67
	épidémies			0	0.0				
	épidémique			1	11.11				
spread	propagation	75	70	53	5732.65	5967.35	5198.22	5569.53	4851.67
	propagé			9	165.31				
	propagée			0	0.0				
	se propagent			0	0.0				
	diffusion			0	0.0				
	se propager			5	51.02				
	se propage			3	18.37				
	propagés			0	0.0				
SARS-CoV	SARS-CoV	73	0	0	0	0	0	0	0
vaccine	vaccinale	68	68	0	0.0	10000.0	10000.0	10000.0	10000.0
	vaccin			68	10000.0				
transmission	transmission	67	64	64	10000.0	10000.0	9124.53	9552.24	8715.97
	se transmettre			0	0.0				
	transmis			0	0.0				
Opus model (En-Fr)									
Src Term	**Tgt Term**	**|t|**	**k**	**f**	$(\frac{f}{k} \times 100)^2$	C_t	**CoTERM**$_{\text{f}}$	C_t^w	**CoTERM**$_{\text{f}}^{\text{w}}$
SARS-CoV-2	SRAS-CoV-2	178	172	152	7809.63	7944.84	7418.26	7677.04	7168.21
	SARS-CoV-2			20	135.21				
virus	virus	164	161	159	9753.1	9754.64	9401.03	9576.2	9229.06
	viral			2	1.54				
	virale			0	0.0				
	virales			0	0.0				
symptoms	symptômes	104	102	102	10000.0	10000.0	9619.08	9807.69	9434.1
outbreak	épidémie	88	75	74	9735.11	9736.89	7072.57	8298.49	6027.76
	épidémies			0	0.0				
	épidémique			1	1.78				
spread	propagation	75	72	57	6267.36	6450.62	5944.89	6192.6	5707.09
	propagé			9	156.25				
	propagée			0	0.0				
	se propagent			1	1.93				
	diffusion			0	0.0				
	se propager			2	7.72				
	se propage			3	17.36				
	propagés			0	0.0				
SARS-CoV	SARS-CoV	73	11	11	10000.0	10000.0	227.06	1506.85	34.21
vaccine	vaccinale	68	68	0	0.0	10000.0	10000.0	10000.0	10000.0
	vaccin			68	10000.0				
transmission	transmission	67	66	66	10000.0	10000.0	9703.72	9850.75	9558.89
	se transmettre			0	0.0				
	transmis			0	0.0				

Table 30: Examples of CoTERM and C_t scores obtained by Transf and Opus on the Tico-19 test set.