A test suite for the evaluation of Portuguese-English machine translation

Mariana Avelino[1], Vivien Macketanz[2],
Eleftherios Avramidis[2], and Sebastian Möller[1,2]

[1] Technische Universität Berlin, Berlin, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
m.amaralgarciaavelino@campus.tu-berlin.de, vivien.macketanz@dfki.de,
eleftherios.avramidis@dfki.de, sebastian.moeller@dfki.de

Abstract. This paper describes the development of the first test suite for the language direction Portuguese-English. Designed for fine-grained linguistic analysis, the test suite comprises 330 test sentences for 66 linguistic phenomena and 14 linguistic categories. Eight different MT systems were compared using quantitative and qualitative methods via the test suite: DeepL, Google Sheets, Google Translator, Microsoft Translator, Reverso, Systran, Yandex and an internally built NMT system trained over 30 hours on 2,5M sentences. It was found that ambiguity, named entity & terminology and verb valency are the categories where MT systems struggle most. Negation, pronouns, subordination, verb tense/aspect/mood and false friends are the categories where MT systems perform best.

Keywords: Machine translation · Evaluation · Portuguese · Test suite

1 Introduction

In an increasingly interconnected world, bridging gaps in communication is ever more important. The value of machine translation (MT) is therefore hard to overstate, and hand-in-hand with a great demand for MT is a demand for tools that can evaluate MT output. After all, only through evaluation can weakness in MT systems be identified and addressed.

This study traces the development of a new test suite for the fine-grained linguistic analysis of the language pair Portuguese-English. Eight different translation systems were evaluated using the new test suite: DeepL, Google Sheets, Google Translate, Microsoft Translator, Reverso, Systran and Yandex, along with an in-house NMT system. Key contributions include:

- The first MT test suite for evaluating Portuguese-English translations on a fine-grained level via 66 phenomena organized in 14 categories.[3]

[3] The test suite has been made publicly available to aid further research: https://github.com/mariana200196/testsuite-pt-br_en

- The corroboration of previous research that ambiguity is one of the most challenging linguistic categories for MT systems to resolve.
- The identification of categories where MT systems perform very well on average (negation, pronouns, subordination, verb tense/aspect/mood and false friends) and very poorly on average (ambiguity, named entity & terminology and verb valency).
- The finding that Reverso significantly outperforms other state-of-the-art MT systems in the translation of multi-word expressions, which include idioms, collocation and verbal multi-word expressions.

The paper is structured as follows: Section 2 summarizes related work in the field of MT evaluation. Section 3 describes this paper's methodological approach, including the experimental setup. Section 4 details the findings. Finally, Section 5 concludes the paper and provides an outlook for future research.

2 Related Work

There is general consensus in the scientific community that the most accurate way to evaluate the quality of MT output is via professional human translators. Unfortunately, this method is not scalable. There is thus a need for more automated evaluation processes which are fast and cost-effective. Over the years, various automatic MT evaluation methodologies have been proposed. Currently, the most widely used method is the BLEU score, as it is quick, language-independent and correlates highly with human evaluation [19]. Unfortunately, this method has known limitations [9, 21]. As Barreiro and Ranchhod [5] explain, while a BLEU score might be useful to those who only need black-box answers to questions such as 'does the system work better today than yesterday?' or 'which MT system performs better?', it cannot provide a transparent diagnosis. In an effort to address these limitations, the use of test suites has been proposed.

A test suite investigates several linguistic phenomena and uses non-generic, manually-devised sentences as test sets. It measures quantitative performance and diagnoses qualitative shortcomings in translation. Test suites thereby deliver fine-grained evaluations of translation quality which help researchers form hypotheses as to why certain errors happen (systematically) and come up with strategies for improving the systems [7]. In recent years, various test suites have been created. Most of these focus on particular phenomena (e.g., [6, 8, 11, 20]) with only very few performing a systematic evaluation of multiple phenomena simultaneously. Macketanz et al. [17, 18] and Avramidis et al. [3, 4] perform a systematic evaluation of more than one hundred phenomena for German-English translation, a practice which we also follow for our chosen language pair. Test suites also differ in the way the MT outputs are evaluated. Some test suites rely on manual labour to check the translations (e.g., [12]) while others provide fixed reference translations. Macketanz et al. [16] propose a semi-automatic evaluation powered by regular expressions and limited human annotation, a method which we adopt for our language pair, too.

3 Method

3.1 Creation of the Test Suite

There are four steps to creating a semi-automatic test suite for a new language direction: (1) producing a paradigm, (2) writing regular expressions, (3) fetching translations and (4) resolving warnings.

Given a chosen language pair, categories and subcategories (referred to as "phenomena" in this paper) should be determined for investigation, for example verb tense (category) and simple past (phenomenon). Then, sentences should be devised to test for the phenomena and an annotator should write rules to control the correctness of machine translations. These rules can be subdivided into "positive regular expressions" and "negative regular expressions". Once the test sentences and regular expressions have been created, the test sentences should be given as input to the MT system(s) and the output fetched. After that, the translations should be fed into the test suite. If the MT output matches a positive regular expression, the translation should be considered correct. If the MT output matches a negative regular expression, the translation should be considered incorrect. If a MT output does not match either a positive or negative regular expression, or if these contradict to each other, the automatic evaluation should produce a "warning" to be manually resolved by the annotator.

For each phenomenon, category and system being tested, the test suite should output an accuracy score:

$$\text{accuracy} = \frac{\text{correct\_translations}}{\text{sum\_of\_test\_items}} \tag{1}$$

To reveal the best system, a one-tailed t-test is performed. All the systems which are not significantly worse than the best system should be grouped together with it in a "first class".

3.2 Limitations of the Method

The accuracy calculation described above is a very intuitive way to assess MT quality. There are, however, some general limitations to keep in mind. For instance, systems that excel at handling a few specific phenomena will be at a disadvantage compared to well-rounded systems, even if the well-rounded systems don't excel at any one phenomenon. Also, a very high score for a phenomenon does not necessarily mean that the MT for that phenomenon has been cracked. Perhaps the difficulty of the test sentences simply needs to be raised to offer a better suited "challenge set" [12].

Referring specifically to our test suite, there are two additional limitations to consider. Firstly, the low number of test sentences per phenomenon can be misleading. As there are only five test sentences per phenomenon, relative differences in accuracy between systems loom larger than absolute differences. For example, if system A translates 3/5 sentences correctly for phenomenon P and system B translates 4/5 sentences correctly for phenomenon P, B's accuracy score for P will be a whole 20% higher (80%) than A's (60%) even though the difference is only one sentence. Secondly, the unequal number of phenomena per category creates bias. Systems which perform well in categories that encompass many phenomena are likely to have their performance scores inflated. For example, systems which translate verbs well are likely to get a higher overall score than those systems which struggle to translate verbs correctly, even if the latter systems perform better at many more categories than the former.

3.3 Experimental Setup

The Portuguese-English test suite described in this paper was created by a Brazilian-born native speaker of Portuguese and English. The test sentences are therefore written in Brazilian Portuguese. The test suite comprises 14 categories, 66 phenomena and 5 test sentences per phenomenon. The categories and phenomena were partly inspired by the categories and phenomena present in existing test suites [4], partly by personal observations of common MT errors and partly by previous research [5, 10].

Table 1. Corpora used for training our NMT system.

Corpus	# sentences	Set
Europarl [14]	2,0M	Training
Global Voices [22]	92,0k	Training
backtranslations	25,9k	Training
Books [22]	1,4k	Training
TED-2013 [22]	0,2M	Validation
Tatoeba [22]	0,2M	Validation

Eight different translation systems were evaluated using the test suite: DeepL (deepL), Google Translate (googl), Microsoft Translator (MS), Reverso (revers), Systran (systr), Yandex (yandx), Google Sheets (gglSh)[4] and a NMT system developed internally (own). The first six systems are commercial systems which came highly recommended in blogs for Portuguese speakers seeking translation services. Given that they are commercial systems, they can be thought of as state-of-the-art. Our own system was developed using the Marian NMT framework[5] [13]. Training was conducted over approximately 30 hours and 2,5M sentences (Table 1). Corpora with sentences from spoken language or newspaper language were preferred to keep the vocabulary of the training set as similar as possible to that of the test set. For the same reason, Brazilian Portuguese corpora were chosen over European Portuguese where possible.

[4] Google Sheets, a spreadsheet program with a cell-wise translation function, was chosen to offer an opportunity for comparison against Google Translate.

[5] https://github.com/marian-nmt/marian-examples/tree/master/training-basics-sentencepiece

4 Findings

4.1 Overall Performance of MT Systems

The average accuracies of each system are shown in Table 2. Micro-average refers to equation 1. Category macro-average calculates the mean in such a way that categories are weighted equally and phenomenon macro-average weights the phenomena equally. Google, Reverso and DeepL are the best performing systems for all three accuracy scores with no significant difference in performance. According to the category macro-average and phenomenon macro-average, Microsoft Translator and Systran are also first-class.

Yandex was the worst performing system, doing worse on average than our system. The poor performance of our system can likely be attributed to insufficient training data. With regards to Yandex, one might speculate that poor performance is partly due to the system interpreting all Portuguese inputs as European Portuguese by default. Brazilian and European Portuguese, though very similar, differ at times in terms of spelling and grammar, so a machine trained to expect European Portuguese might struggle when confronted with Brazilian Portuguese. In fact, several Brazilian researches have commented on how training a system on European Portuguese corpora to then translate Brazilian test sentences reduces the BLEU scores of the output [1, 10, 15]. Unlike Yandex, the other commercial systems can distinguish between European and Brazilian Portuguese or default to Brazilian Portuguese.

4.2 BLEU vs. Test Suite Scores

Different studies [4, 11, 12] are divided as to whether system ranking according to BLEU scores correlates with system ranking according to test suite scores. To examine this, reference sentences for the test items were created and a BLEU score was calculated for each system.

The BLEU ranking shuffles the order of the top performing systems (Google, Reverso, DeepL, SYSTRAN, and Microsoft Translator), but not the order of the worst performing ones (Google Sheets, own, and Yandex). Worth noting is that the score gap between the top performing and worst performing systems is far less pronounced in BLEU (only 2 points difference between Microsoft Translator and Google Sheets). After reading all the translated output from the different systems, it becomes evident that the 17 point gap between Microsoft Translator and Google Sheets produced by the test suite is more representative of reality. The BLEU score makes it seem as though the difference in MT quality between Reverso and Microsoft is comparable to the quality difference between Microsoft and Google Sheets when it is not. Microsoft MT quality is far closer to that of Reverso than Google Sheets is to that of Microsoft Translator.

4.3 Categories

The test suite revealed that the best performing category was negation, where all systems scored 100%. Other categories with an average accuracy of 80% or

Table 2. Test suite accuracy (%) per category for 8 Portuguese to English MT systems.

category	#	googl	rever	deepL	MS	systr	gglSh	own	yandx	avg
ambiguity	11	54.5	72.7	72.7	63.6	54.5	45.5	$^{\downarrow}$27.3	54.5	55.7
coord. & ellipsis	18	**100.0**	**100.0**	**88.9**	77.8	**88.9**	38.9	61.1	44.4	75.0
false friends	5	80.0	80.0	80.0	80.0	80.0	80.0	100.0	60.0	80.0
function word	13	76.9	76.9	76.9	76.9	76.9	46.2	30.8	38.5	62.5
ldd & interrogative	50	**82.0**	**74.0**	**70.0**	**70.0**	**72.0**	48.0	44.0	56.0	64.5
mwe	19	**57.9**	**$^{\uparrow}$84.2**	**68.4**	**63.2**	**63.2**	**68.4**	47.4	**63.2**	64.5
ne & terminology	20	**75.0**	**50.0**	**65.0**	**60.0**	**65.0**	**55.0**	$^{\downarrow}$30.0	**50.0**	56.3
negation	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
non-verbal agreement	5	**80.0**	**100.0**	**80.0**	**80.0**	**40.0**	**40.0**	**60.0**	0.0	60.0
pronouns	13	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	61.5	76.9	38.5	84.6
punctuation	10	80.0	60.0	50.0	70.0	70.0	60.0	60.0	70.0	65.0
subordination	39	**92.3**	**92.3**	**92.3**	**89.7**	**89.7**	74.4	74.4	66.7	84.0
v. tense/aspect/mood	113	**92.0**	**92.0**	**92.9**	**90.3**	**91.2**	73.5	69.0	60.2	82.6
verb valency	10	80.0	60.0	80.0	60.0	50.0	60.0	30.0	40.0	57.5
categ. macro-average	331	**82.2**	**81.6**	**79.8**	**77.3**	74.4	60.8	57.9	53.0	70.9
phen. macro-average	331	**85.1**	**83.7**	**82.5**	79.7	**80.5**	62.4	57.7	55.8	73.4
micro-average	331	**85.5**	**84.0**	**83.1**	**80.4**	**80.7**	63.1	58.6	56.5	74.0
BLEU	331	54.3	49.5	54.8	47.2	51.5	45.1	38.0	27.7	46.0

Boldface indicates the best accuracies in every category (row) based on a one-tailed t-test. Accuracies two standard deviations higher and lower than the average per category are indicated respectively by $^{\uparrow}$ and $^{\downarrow}$. Test sentences which produced warnings are excluded from the accuracy calculations.

more were pronouns (84,6%), subordination (84,0%), verb tense/ aspect/ mood (82,6%) and false friends (80%).

The worst performing category was ambiguity with an average score of 55,7%. Studies into MT quality for English-Portuguese [5,10] likewise found that translation errors relating to ambiguous words were among the most common. Other categories where the systems performed poorly (below 60%) were named entity & terminology (56,3%) and verb valency (57,5%).

In Table 2 the systems which performed more than two standard deviations above the mean and those which performed more than two standard deviations below the mean are indicated with upward-facing or downward-facing arrows, respectively. Reverso performed extremely well at translating multi-word expressions (MWE) in comparison to other systems. This category encompasses phenomena such as idioms. Our system performed quite poorly in the categories ambiguity and named entity & terminology. Its poor performance handling ambiguity likely correlates with previous findings that NMT systems require a far higher amount of training data to learn how to translate ambiguous words correctly relative to other phenomena [2,10]. On a related note, our system probably lacked sufficient exposure to location names, proper names etc. in the training data, and so failed to correctly translate many named entities.

Table 3. Test suite accuracy (%) of the 10 worst performing phenomena.

phenomenon	#	googl	rever	deepL	MS	systr	gglSh	own	yandx	avg
direct object omissions & polar questions	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
idiom	4	0.0	$^{\uparrow}$**75.0**	**25.0**	0.0	**25.0**	**25.0**	0.0	0.0	18.8
indicativo pretér. imperf.	5	**40.0**	**40.0**	**40.0**	**40.0**	0.0	**20.0**	0.0	**20.0**	25.0
proper name	5	**60.0**	**20.0**	**40.0**	**20.0**	**40.0**	0.0	**20.0**	**20.0**	27.5
quotation marks	5	**60.0**	**20.0**	0.0	**40.0**	**40.0**	**20.0**	**40.0**	**40.0**	32.5
mediopassive voice	5	**60.0**	**40.0**	**60.0**	**40.0**	**40.0**	**40.0**	0.0	0.0	35.0
focus particle	5	40.0	40.0	40.0	40.0	40.0	60.0	20.0	40.0	40.0
domain specific	5	**40.0**	**40.0**	**40.0**	**40.0**	**40.0**	**60.0**	$^{\downarrow}$0.0	**60.0**	40.0
null object	3	**100.0**	**100.0**	**66.7**	**66.7**	**100.0**	0.0	0.0	**33.3**	58.3
collocation	5	20.0	60.0	40.0	40.0	20.0	60.0	40.0	60.0	42.5

Boldface indicates the best accuracies in every phenomenon (row) based on a one-tailed t-test. Accuracies two standard deviations higher and lower than the average per phenomenon are indicated respectively by $^{\uparrow}$ and $^{\downarrow}$. Test sentences which produced warnings are excluded from the accuracy calculations.

4.4 Phenomena

Table 3 shows the ten most incorrectly translated phenomena (the full table can be found in the project github). A comparison reveals that some of the phenomena with the most translation errors on average (e.g., proper name, mediopassive voice) indeed belong to some of the worst performing categories, but not all. For example, the phenomenon direct object omission & polar question (which did not have a single sentence translated correctly) belongs to the category coordination & ellipsis, which is not among the categories with the most translation errors. Furthermore, indicativo pretérito imperfeito is also in the bottom 10 phenomena, yet it belongs to one of the best performing categories: verb tense/aspect/mood. Were the test suite less fine-grained, some problematic phenomena would have remained hidden within well-performing categories.

Reverso is the only system that performed two standard deviations better than the mean, doing so for idioms. Idioms belong to the category MWE, where Reverso also achieved an accuracy two standard deviations above average. The 3 worst performing MT systems overall (Google Sheets, own and Yandex) had accuracies two standard deviations below the mean for multiple phenomena.

4.5 Qualitative Analysis

By allowing the inspection of test sentences and their translations, test suites additionally help researchers understand where MT systems are struggling and why. Here we examine 4 phenomena to develop assumptions about their errors.

Mediopassive Voice Mediopassive voice asserts that a person or thing both performs and is affected by the action represented. A Portuguese example is

Table 4. Examples of phenomena with failing and (if existing) passing MT outputs.

Mediopassive Voice	
Vendem-se casas.	
Houses are for sale.	*reference translation*
Houses sell.	fail
This house is for sale.	pass
Direct Object Omission & Polar Question	
Ele estuda todos os dias? Estuda.	
Does he study everyday? Yes./ He does./ Yes, he does.	*reference translation*
Does he study everyday? Studies.	fail
Idiom	
Está chovendo a cântaros.	
It's raining cats and dogs./ It's raining heavily.	*reference translation*
It's raining vases.	fail
It's raining cats and dogs.	pass
False Friends	
Onde você pôs a agenda da vice-diretora?	
Where did you put the deputy director's planner?	*reference translation*
Where have you put the deputy director's agenda?	fail

presented in Table 4. The incorrect, literal translation of "Vendem-se casas" was not an isolated incident. An examination of all test sentences revealed that the systems tended to translate mediopassive voice word-for-word. This inevitably produced wrong outputs, because mediopassive sentences in Portuguese must generally be converted into passive or active voice to preserve their meaning in English. This complexity is compounded by the rarity of mediopassive voice, making it a challenging phenomenon indeed for MT systems.

Direct Object Omission & Polar Question In Portuguese, when replying to a 'yes' or 'no' question (polar question), it is uncommon to answer with 'yes' or 'no'. Instead, the verb from the question is used as a one-word reply and any direct/indirect object is omitted. The example test sentence in Table 4 has a very straightforward translation, yet all systems failed completely. A common output resulted from literally translating "Estuda" into "Studies". After inspecting the incorrect translations, one might hypothesize that the systems' widespread failure is due to their insensitivity to inter-sentence context.[6]

Idiom An idiom is a group of words established by usage that have a meaning not deducible from those of the individual words. They present a challenge to human translators and machines alike because the figurative nature of idioms usually demands interpretation and explanation during translation. An analysis of the MT outputs revealed that systems often successfully translated Portuguese idioms which had an equivalent English idiom (see Table 4). In contrast, idioms which did not have an English equivalent were consistently mistranslated.

[6] During the paper review process, the test sentences for direct object omissions & polar question were re-translated and DeepL translated all of them correctly.

False Friends A false friend is a set of words that in different languages look or sound similar, but differ in meaning. There is the expectation that machines should not mistranslate false friends because they "learn" only what words in one language map to in the other language. Machines should therefore be impervious to the cues that mislead humans, namely how a word sounds and looks. While mistranslations are rare, Table 4 reveals that they can still happen in exceptional cases when a word is both a false friend (e.g. "agenda" is a word in Portuguese and English) and lexically ambiguous ("agenda" in Portuguese can mean either "planner" (a journal) or "agenda" (someone's underlying plan).[7] The test suite has found MT systems to be more robust against false friends than lexical ambiguity, so it is likely that what was classified as a false friend error is in fact a consequence of lexical ambiguity, but we cannot be certain.

5 Conclusion

As part of this research, the first test suite for the language direction Portuguese-English was developed. It is designed for fine-grained linguistic analysis and comprises 330 test sentences for 66 phenomena and 14 categories. Via the test suite, the translation quality of eight MT systems was evaluated quantitatively and qualitatively (DeepL, Google Sheets, Google Translator, Microsoft Translator, Reverso, Systran, Yandex and our own system). It was found that ambiguity remains one of the most challenging linguistic categories for MT systems. Alongside ambiguity, named entity & terminology and verb valency are the categories where MT systems fail the most on average. On a phenomenon-level, direct object omissions & polar questions is where all systems struggled the most. Positive findings were that negation, pronouns, subordination, verb tense/aspect/mood and false friends are the categories where MT systems perform the best on average. It was also observed that Reverso performs exceptionally well in the translation of multi-word expressions, in particular idioms. In order to aid future research, this test suite has been made publicly available.

We see three main areas for improvement: (1) increasing the number of test sentences per phenomena to allow for more statistically sound and reliable observations, (2) developing a complementary English-Portuguese test suite and (3) enriching the test suite with harder test sentences, as well as new phenomena.

Acknowledgments

The work was accomplished as a semester project, co-ordinated by Neslihan Iskender, part of the MSc program of Media Informatics (TU Berlin). The supervision was funded by the projects TextQ (German Research Foundation; DFG) and SocialWear (German Ministry of Research and Education; BMBF).

[7] The lexical ambiguity of "agenda" was overlooked when the test sentence was created; test sentences should only test one phenomenon at a time. More philosophically-minded readers might therefore want to debate whether the incorrect translation boils down to a human or machine error.

References

1. Aires, J., Lopes, G., Gomes, L.: English-Portuguese biomedical translation task using a genuine phrase-based statistical machine translation approach. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp. 456–462. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2335
2. Avramidis, E., Macketanz, V., Lommel, A., Uszkoreit, H.: Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In: Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing. pp. 243–248. Association for Machine Translation in the Americas, Boston, MA (Mar 2018), https://www.aclweb.org/anthology/W18-2107
3. Avramidis, E., Macketanz, V., Strohriegel, U., Burchardt, A., Möller, S.: Fine-grained linguistic evaluation for state-of-the-art machine translation. In: Proceedings of the Fifth Conference on Machine Translation. pp. 346–356. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.wmt-1.38
4. Avramidis, E., Macketanz, V., Strohriegel, U., Uszkoreit, H.: Linguistic evaluation of German-English machine translation using a test suite. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 445–454. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-5351
5. Barreiro, A., Renchhod, E.: Machine translation challenges for portuguese. Lingvisticæ Investigationes **28**, 3–18 (2005). https://doi.org/10.1075/li.28.1.03bar
6. Bojar, O., Mírovský, J., Rysová, K., Rysová, M.: EvalD reference-less discourse evaluation for WMT18. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 541–545. Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6432
7. Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.T., Williams, P.: A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. The Prague Bulletin of Mathematical Linguistics **108**, 159–170 (2017). https://doi.org/10.1515/pralin-2017-0017
8. Burlot, F., Scherrer, Y., Ravishankar, V., Bojar, O., Grönroos, S.A., Koponen, M., Nieminen, T., Yvon, F.: The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 546–560. Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6433
9. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the Role of BLEU in Machine Translation Research. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. pp. 249–256. Trento, Italy (Apr 2006). https://doi.org/10.1145/1083784.1083789
10. Caseli, H., Inácio, M.: NMT and PBSMT error analyses in English to Brazilian Portuguese automatic translations. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 3623–3629. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.446
11. Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., Loáiciga, S.: A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 570–577.

Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6435

12. Isabelle, P., Cherry, C., Foster, G.: A challenge set approach to evaluating machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2486–2496. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1263
13. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++ pp. 116–121 (Jul 2018). https://doi.org/10.18653/v1/P18-4020
14. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the tenth Machine Translation Summit. vol. 5, pp. 79–86. Phuket, Thailand (2005), http://mt-archive.info/MTS-2005-Koehn.pdf
15. Lopes, A., Nogueira, R., Lotufo, R., Pedrini, H.: Lite training strategies for Portuguese-English and English-Portuguese translation. In: Proceedings of the Fifth Conference on Machine Translation. pp. 833–840. Association for Computational Linguistics, Online (2020), https://aclanthology.org/2020.wmt-1.90
16. Macketanz, V., Ai, R., Burchardt, A., Uszkoreit, H.: TQ-AutoTest – an automated test suite for (machine) translation quality. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1142
17. Macketanz, V., Avramidis, E., Burchardt, A., Uszkoreit, H.: Fine-grained evaluation of German-English machine translation based on a test suite. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 578–587. Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6436, https://aclanthology.org/W18-6436
18. Macketanz, V., Avramidis, E., Manakhimova, S., Möller, S.: Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In: Proceedings of the Sixth Conference on Machine Translation. pp. 1059–1073. Association for Computational Linguistics, Online (Nov 2021), https://aclanthology.org/2021.wmt-1.115
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. Association for Computational Linguistics, USA (2002). https://doi.org/10.3115/1073083.1073135
20. Rios, A., Müller, M., Sennrich, R.: The word sense disambiguation test suite at WMT18. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 588–596. Association for Computational Linguistics, Belgium, Brussels (Oct 2018). https://doi.org/10.18653/v1/W18-6437
21. Smith, A., Hardmeier, C., Tiedemann, J.: Climbing mont BLEU: The strange world of reachable high-BLEU translations. In: Proceedings of the 19th Annual Conference of the European Association for Machine Translation. pp. 269–281 (2016), https://aclanthology.org/W16-3414
22. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf