cushLEPOR: Customised hLEPOR Metric Using LABSE Distilled Knowledge Model to Improve Agreement with Human Judgements

Lifeng Han[1], **Irina Sorokina**[2], **Gleb Erofeev**[2], and **Serge Gladkoff**[2]
[1] ADAPT Research Centre, DCU, Ireland
[2] Logrus Global, Translation & Localization
gleberof, irina.sorokina, serge.gladkoff@logrusglobal.com

Abstract

Human evaluation has always been expensive while researchers struggle to trust the automatic metrics. To address this, we propose to customise traditional metrics by taking advantages of the pre-trained language models (PLMs) and the limited available human labelled scores. We first re-introduce the hLEPOR metric factors, followed by the Python portable version we developed which achieved the automatic tuning of the weighting parameters in hLEPOR metric. Then we present the customised hLEPOR (cushLEPOR) which uses LABSE distilled knowledge model to improve the metric agreement with human judgements by automatically optimised factor weights regarding the exact MT language pairs that cushLEPOR is deployed to. We also optimise cushLEPOR towards human evaluation data based on MQM and pSQM framework on English-German and Chinese-English language pairs. The experimental investigations show cushLEPOR boosts hLEPOR performances towards better agreements to PLMs like LABSE with much lower cost, and better agreements to human evaluations including MQM and pSQM scores, and yields much better performances than BLEU (data available at https://github.com/poethan/cushLEPOR).

1 Introduction

Machine Translation (MT) is a rapidly developing research field that plays an important role in NLP area. MT started from 1950s as one of the earliest artificial intelligence (AI) research topics and gained a large improvement in the output quality in large resourced language pairs after the introduction of Neural MT (NMT) in recent years (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014). However, the challenge still remains in achieving human parity MT output (Han et al., 2021a). Thus MTE continues to play an important role in aiding MT development from the aspects of timely and high quality evaluations, as well as reflecting the translation errors that MT systems can take advantages of for further improvement (Han et al., 2021b). On the one hand, human evaluations have long been criticised as expensive and unrepeatable. Furthermore, the inter and intra-agreement levels from Human raters always struggle to achieve a relatively high and reliable score. On the other hand, automatic evaluation metrics have reached high performances in the category of system level evaluations of MT systems (Freitag et al., 2021). However, the segment level performance is still a large gap from human experts' expectation.

In the meantime, many pre-trained language models have been proposed and developed in very recent years and showing big advantages in different NLP tasks, for instance, BERT (Devlin et al., 2019) and its further developed variants (Feng et al., 2020). In this work, we take the advantages of both high performing automatic metric and pre-trained language model, aiming at one step further towards higher quality performing automatic MT evaluation metric from both system level and segment level perspectives.

Among the evaluation metrics developed recent years, hLEPOR (Han et al., 2013b,a) is one of the comprehensive category that include many evaluation factors including precision, recall, word order (via position difference factor), and sentence length. It has also been applied by many researchers from different NLP field including natural language generation (NLG) (Novikova et al., 2017; Gehrmann et al., 2021), natural language understanding (NLU) (Ruder et al., 2021), automatic text summarization (ATS) (Bhandari et al., 2020), and searching (Liu et al., 2021), in addition to MT evaluation (Marzouk, 2021).

However, there are some disadvantages from hLEPOR including the manual tuning of its parameter weights which costs lots of human efforts. We choose hLEPOR (Han et al., 2013b,a) as our baseline model, and use the very recent language model LABSE (Feng et al., 2020) to achieve automatic tuning of its parameters thus aiming at reducing the evaluation cost and further boosting the performance. This system description paper is based on our earlier work, especially the training models (Erofeev et al., 2021).

The rest of the paper is organised as bellow: Section 2 revisits hLEPOR metric, its factors, advantages and disadvantages, Section 3 introduces our Python ported version of hLEPOR and the further customised hLEPOR (cushLEPOR) using language models, Section 4 presents our experimental development and evaluation that we carried out on cushLEPOR metric using WMT historical data, Section 5 reserves space for our submission to this year WMT21 metrics task, and Section 6 finishes this paper with discussions of our findings and possible future work.

2 Revisiting hLEPOR

hLEPOR is a further developed variant of LEPOR (Han et al., 2012) metric which was firstly proposed in 2013 including all evaluation factors from LEPOR but using harmonic mean for grouping factors to produce final calculation score (Han et al., 2013b). Its submission to WMT2013 metrics task achieved system level highest average correlating scores to human judgement on English-to-other (French, Spanish, Russian, German, Czech) language pairs by Pearson correlation coefficient (0.854) (Han, 2014; Macháček and Bojar, 2013). Other MT researchers also analysed LEPOR metric variant as one of the best performing segment level metric that was not significantly outperformed by other metrics using WMT shared task data (Graham et al., 2015). hLEPOR is calculated by:

$$h\text{LEPOR} = Harmonic(w_{LP}LP, \\ w_{NPosPenal}NPosPenal, w_{HPR}HPR)$$

where *LP* is a sentence length penalty factor which was extended from briefty penalty utilised in BLEU metric, *NPosPenal* is for n-gram position difference penalty which captures the word order information, and *HPR* is the harmonic mean of Precision and Recall values. We refer the work (Han, 2014) for detailed factor calculation with examples there.

The basic version of hLEPOR carries out similarity calculation between MT system outputs and reference translations, in the same language setting, based on the *word surface level* tokens. The hybrid hLEPOR metric also carries out similarity calculation based on POS sequences from system-output and reference text. To do this, POS tagging is needed as the first step, then hLEPOR(POS) calcualation uses the same algrithms used for the word level similarity score hLEPOR(word). Finally, hybrid hLEPOR is a combination of both word level and POS level score. In this system submission work, with the time limitations, to make an easier to use customised hLEPOR, we take the basic version of hLEPOR, i.e. the word level simialrity calculation and leave the hybrid hLEPOR into the future work.

The weighting parameters for the three main factors in original hLEPOR metric, i.e the (w_{LP}, $w_{NPosPenal}$, w_{HPR}) set, in addition to the other parameters inside each factor, was tuned by manual work based on development data. This is very time consuming, tedious, and costly. In this work, we will introduce an automated tuning model for hLEPOR to customise it regarding deployed language pairs, which we name as cushLEPOR.

3 Proposed Model

3.1 Python portable hLEPOR

Original hLEPOR was published as Perl code [1], in a non-portable format, which is not very suitable for modern AI/NLP applications, since they are using almost exclusively Python. Python is a programming language of choice for AI/ML tasks, thanks to its amazing ecosystem of open source or simply free libraries available to researchers and developers. However, hLEPOR was not available in NLTK (Bird et al., 2009) or any other public Python libraries. We therefore took original published Perl code and ported it to Python, carefully comparing the logic of original paper and the Perl implementation. During this work we run both Perl code to reproduce the results of original code, and the new Python implementation. This work helped us to spot and fix at least three minor errors which did not significantly affected the score, but nevertheless we fixed the bugs of the Perl code.

While doing the porting we did also notice that hLEPOR parameter values were taken empirically and never explained in detailed except for the suggested parameter setting table in the paper (Han et al., 2013b,a) for eight language pairs that were

[1] https://github.com/poethan/LEPOR

tested for the WMT2013 shared task, including EN-CZ/DE/FR/ES and the opposite direction. They were:

- **alpha**: the tunable weight for recall
- **beta**: the tunable weight for precision
- **n**: words count before and after matched word in npd calculation
- **weight_elp**: tunable weight of enhanced length penalty
- **weight_pos**: tunable weight of n-gram position difference penalty
- **weight_pr**: tunable weight of harmonic mean of precision and recall

The parameter values for hLEPOR as published in the publicly available Perl code were (from *CZ-EN language pair* setting (Han et al., 2013b,a)):

- **alpha** = 9.0,
- **beta** = 1.0,
- **n** = 2,
- **weight_elp** = 2.0,
- **weight_pos** = 1.0,
- **weight_pr** = 7.0.

We came to the conclusion that we need to check whether these parameters are optimal, and find out whether better set of values exist to improve agreement with human judgement.

Because the different characteristics of each language, and language families, the evaluation of MT outputs would emphasis on different factors. For instance, word order factor reflected by n-gram position different penalty in hLEPOR (NPosPenal), can be with higher or lower weight for strict order languages and loose/flexible word order languages. Thus, we assumed that hLEPOR optimisation towards different languages will generate corresponding different set of parameter values. We call this step language-specific optimisation, and it will save much cost and time to achieve an automatic tuning process. The Python ported hLEPOR is available at Pypi `https://pypi.org/project/hLepor/`.

3.2 cushLEPOR: customised hLEPOR

With the recent development of pre-trained neural language models and their effective applications in different NLP tasks, including question answering, language inference and MT, it becomes a natural question that why do not we apply them in MT evaluation as well.

Very recent work from Google team verified that MQM (Multi-dimension quality metric) (Lommel et al., 2014) and SQM (Scalar Quality Metrics) (Freitag et al., 2021) have good agreement with each other when they were carried out both by professional translators. However, this does not correlate to Mechanical Turk based crowd-sourced human evaluation that was carried out by general researchers or untrained online workers. It also reflected that crowd-sourced evaluation tends to favour very literal translations instead of better translations with more diverse meaning equivalent lexical choices.

To customise hLEPOR (cushLEPOR) towards better parameter setting using pre-trained language models (LMs) for deployed language pairs, we choose Optuna open source hyper-parameter optimisation framework (Akiba et al., 2019) to automate hyper-parameter search for best agreement between cushLEPOR and human experts evaluation.

SQM (Freitag et al., 2021) borrows WMT shared task settings to collect segment-level scalar rating, but set the score scale from 0 to 6 instead of 0 to 100. Professional translator labelled scores using SQM is named as pSQM.

We aim at optimising cushLEPOR parameters to obtain best agreement with pSQM scores. However, in real life, human evaluations are not often feasible to obtain due to the constrains from both time and financial aspects.

We therefore propose to carry out an alternative optimisation model, i.e. customising cushLEPOR parameters towards LABSE (Language Agnostic BERT Sentence Embedding) model similarity score.

LABSE model is built on BERT (Bidirectional Encoder Representations from Transformers) architecture and trained on monolingual (for dictionaries) and bilingual training data. LABSE training data is filtered and processed. The resulting sentence embeddings achieve excellent performance on measures of sentence embedding quality such as the semantic textual similarity (STS) benchmark and sentence embedding based transfer learning (Feng et al., 2020).

LABSE linguistic similarity score finds matching translations very well. The disadvantages, however, are high demand for computational resources (with GPU), intensive application coding and slow performance.

The design of using optimised hLEPOR (cushLEPOR) in lieu of LABSE similarity aims at developing a simple, high-performing, easy to run and not computationally demanding script to achieve results similar to high-end LABSE similarity score, and hopefully towards human judgement. cushLEPOR parameters can be optimised for agreement with any type of score, such as pSQM, MQM, and LABSE, etc.

4 Experimental Evaluations

The training and development data we used regarding MQM scores and pSQM labels is from the recent work by Google Research team on investigating into human evaluations based on WMT2020 shared task (Freitag et al., 2021) (data available at `https://github.com/google/wmt-mqm-human-evaluation`).

We first focus on English-to-German language (EN-DE) pair, which includes MQM and pSQM labels, acquired from 10 submission of WMT 2020, then take ZH-EN dataset. We refer to the paper (Freitag et al., 2021) for detailed MT system names and offering institutions.

Firstly, a multi-parameter optimisation against LABSE for EN-DE language pair gave the following values for cushLEPOR parameters:

- **alpha** = 2.97,
- **beta** = 1.97,
- **n** = 4,
- **weight_elp** = 1.0,
- **weight_pos** = 14.97,
- **weight_pr** = 2.2.

This set of values reflected very different weighting systems in comparison to the original hLEPOR metric. For instance, 1) cushLEPOR assigned recall and precision much closer weight (2.97 vs 1.97) in comparison to hLEPOR (9.0 vs 1.0), 2) cushLEPOR chose 4-gram in chunk matching instead of bi-gram used in hLEPOR, 3) cushLEPOR assigned NPosPenal (n-gram position difference penalty) factor a very heavy weight against other two factors LP (length penalty) and HPR (harmonic mean of precision and recall) by (**14.97** vs 1.0 and 2.2) in comparison to hLEPOR which emphasised the weight on HPR (1.0 vs 2.0 and **7.0**). From these points of view, cushLEPOR trained on EN-DE language pair indicates the importance of the larger window context consideration during word matching, as well as the word order information reflected by n-gram (n value) and novel factor NPosPenal introduced by hLEPOR respectively.

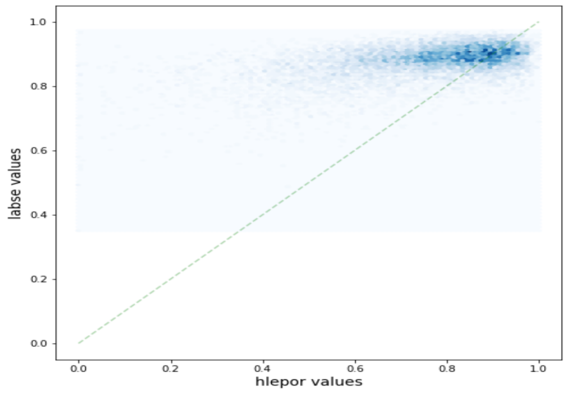

Figure 1: Agreement with LABSE: hLEPOR

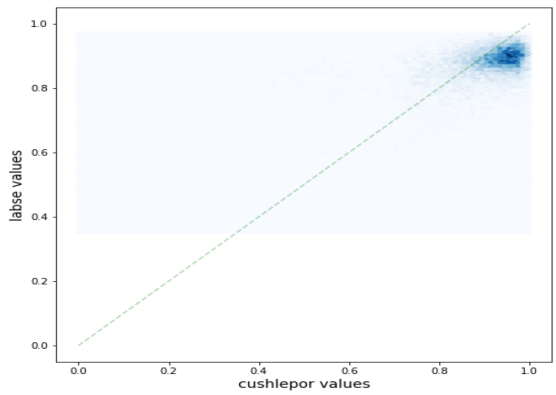

Figure 2: Agreement with LABSE: cushLEPOR

This also reflected that LABSE similarity is indeed a feasible goal for cushLEPOR optimisation. The correlations of hLEPOR and cushLEPOR to LABSE are shown in Fig. 1 and 2.

However, we found out that we were not able to decrease much on RMSE (Root Mean Square Error) score for cushLEPOR towards pSQM , in comparison to original hLEPOR, (0.28 vs 0.29)which does indicate that original hLEPOR empirically shows very good fit for pSQM type human evaluation, using the suggested parameter settings for

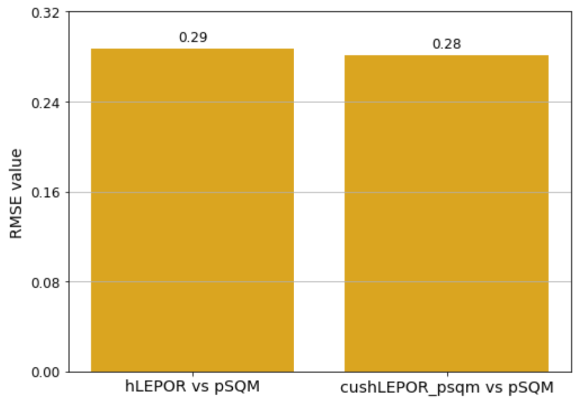

Figure 3: RMSE: hLEPOR vs cushLEPOR to pSQM

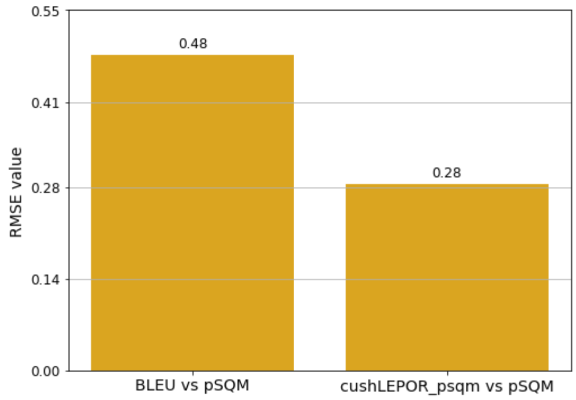

Figure 4: RMSE: BLEU vs cushLEPOR to pSQM

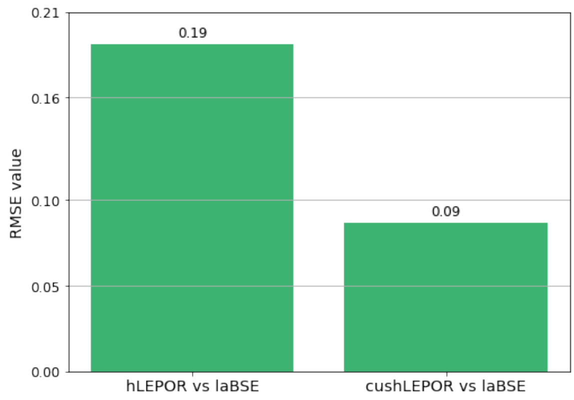

Figure 5: RMSE: hLEPOR vs cushLEPOR to LABSE

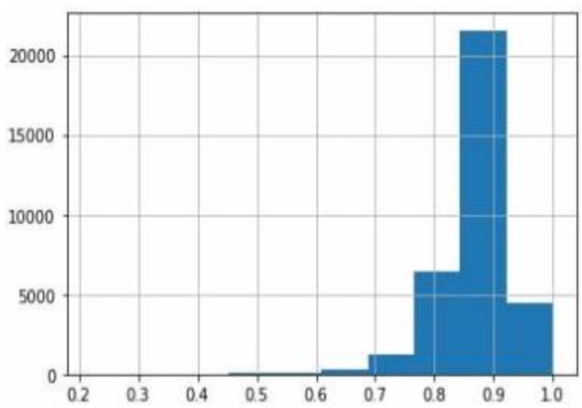

Figure 6: Score Distribution: tune on LABSE

EN-DE (Han et al., 2013a,b) as bellow.

- **alpha** = 9.0,
- **beta** = 1.0,
- **n** = 2,
- **weight_elp** = 3.0,
- **weight_pos** = 7.0,
- **weight_pr** = 1.0.

The RMSE value between pSQM and hLEPOR, vs pSQM and cushLEPOR is shown in Fig. 3. However, it indeed shows much better performance then BLEU metric, as in Fig. 4 (0.28 vs 0.46).

Optuna did optimise cushLEPOR against LABSE very well, halving the RMSE distance between LABSE and cushLEPOR as compared to original hLEPOR, shown in Fig. 5.

The performances of tuning on LABSE and pSQM are shown in Fig. 6 and 7 respectively. The horizontal axis is the score value (0, 1) and the vertical axis is the sentence number that falls into the corresponding score intervals.

From the score distribution visualisation, it reflects the tuning on pSQM has a larger covered error types while LABSE is less sensitive to some errors that human experts would spot out. As shown on these charts, pSQM human rating shows much wider "tail" of "low score ratings", while LABSE rating is much more focused. The reason is that LABSE similarity model underestimates the severity of errors and error types, while humans analyse the meaning and assign proper error penalties in more diverse setting. As an example, the sentence "The comet did not struck the Earth this time." and "The comet did struck the Earth this time." has very close lexical similarity, but the meaning is very different, in this case “opposite”. LABSE similarity score would not assign significant penalty to such difference, while human will treat it as a major error. This difference plays a crucial role for reliable translation quality evaluation.

5 Submission to WMT21

For WMT2021 Metrics Task, we submitted our system scores for zh-en and en-de language pairs, both

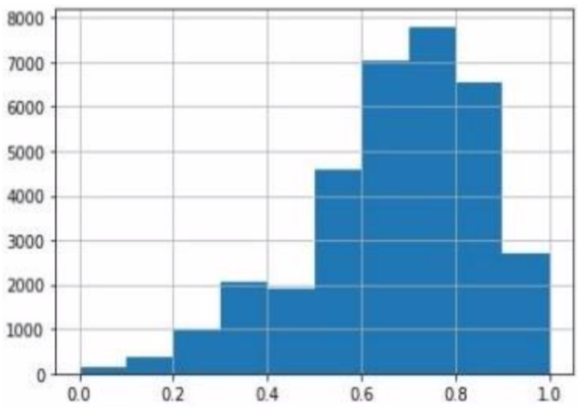

Figure 7: Score Distribution: tune on pSQM

segment-level and system-level evaluation. The training and development set we used are exact the ones from last section (Section 4). We can not tune our model parameters on en-ru language pair from the WMT21 official data, because the human labelled MQM and pSQM scores as validation data that cushLEPOR requires do not exist from last year WMT20 set. WE carried out evaluation on all four official data-sets: **newstest2021** (traditional task), **florestest2021** (sentences translated as part of the WMT News translation task), **tedtalks** (additional sets of sentences translated by WMT21 translation systems in the TED talks domain), and **challengeset** (synthetic outputs generated specifically to challenge automatic metrics).

The optimised parameter values set for our zh-en submission to WMT21 is displayed below:

For cushLEPOR(LM) using LaBSE training:

- **alpha** = 2.85,
- **beta** = 4.73,
- **n** = 1,
- **weight_elp** = 1.01,
- **weight_pos** = 11.13,
- **weight_pr** = 4.62

For cushLEPOR(pSQM) using professional translator labelled SQM training:

- **alpha** = 9.09,
- **beta** = 3.55,
- **n** = 3,
- **weight_elp** = 1.01,
- **weight_pos** = 14.98,
- **weight_pr** = 1.57

The optimised parameter values set for our en-de submission to WMT21 is displayed below:

For cushLEPOR(LM) using LaBSE training:

- **alpha** = 2.95,
- **beta** = 2.68,
- **n** = 2,
- **weight_elp** = 1.0,
- **weight_pos** = 11.79,
- **weight_pr** = 1.87

For cushLEPOR(pSQM) using professional translator labelled SQM training:

- **alpha** = 1.13,
- **beta** = 1.71,
- **n** = 2,
- **weight_elp** = 1.06,
- **weight_pos** = 11.90,
- **weight_pr** = 1.01

6 Discussions and Future Work

In this work, we described cushLEPOR, a customised hLEPOR metric which can be automatically trained and optimised using both human labelled MQM and pSQM scores, as well as large scale pre-trained language model (LM) LABSE towards better agreement to human experts level judgements and distilled LM performance respectively, and reducing cost at the meantime, e.g. the manual tuning from hLEPOR and high computational demand from LMs.

We also optimised cushLEPOR towards human translators' evaluation scores, i.e. pSQM, which showed much improved performance than BLEU and original hLEPOR (with default parameters). Our research is in line with the MT evaluation guideline suggestions from the very recent work (Marie et al., 2021) that better evaluation metrics in correlation to human judgement shall be tested and deployed. Or human judgements shall be carried out directly wherever possible.

We have some findings during the experimental investigation: 1) cushLEPOR trained on LABSE can replace LABSE to carry out similarity calculation task in MT evaluation, which is much more light weighted and low cost from computational power and complexity point of view. 2) we can choose alternative pre-trained language models (LMs) in the future to boost performance. 3) this cushLEPOR optimisation framework proves to be functional, offering high performance towards pre-trained LMs, much improved agreement of cushLEPOR to LABSE scores in comparison to hLEPOR (as in Figure 1 and 2). 4) optimised cushLEPOR achieves better agreement towards professional translator's evaluation (pSQM).

Optuna can generate different set of cushLEPOR parameter values in different runs, which could be an consistency issue. However, we believe it optimises the performance of cushLEPOR towards the highest agreement to the reference scoring (pre-trained LMs or human evaluations), but not to ensure the same set of parameter values to be generated, so this will not be an issue. We will carry out further analysis on this aspect in the future work.

The hybrid version of hLEPOR (Han et al., 2013b) use POS features to function as pseudo synonyms to capture alternative correct translations. However it relays on POS taggers for target language, which does not exist for newly proposed languages, and its tagging accuracy may be low, and it cost extra processing steps. In the future work, we plan to carry out integrated model which combine the POS tagging as a command function in data pre-processing for hybrid cushLEPOR.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. We thank the following open-source teams: Google Research for sharing their human evaluation data on MQM and pSQM scores, Optuna open-source hyper-parameter optimisation framework, and LaBSE pre-trained LMs.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. O'Reilly Media Inc.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gleb Erofeev, Irina Sorokina, Lifeng Han, and Serge Gladkoff. 2021. cushlepor uses labse distilled knowledge to improve correlation with human translation evaluations. In *Proceedings for the Machine Translation Summit - User Track (In Press)*, online. Association for Computational Linguistics & AMTA.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *arXiv e-prints*, page arXiv:2104.14478.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi

Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. *arXiv e-prints*, page arXiv:2102.01672.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1183–1191.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India. The COLING 2012 Organizing Committee.

Aaron Li-Feng Han, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, Yiming Wang, and Jiaji Zhou. 2013a. A description of tunable machine translation evaluation systems in WMT13 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 414–421, Sofia, Bulgaria. Association for Computational Linguistics.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau, MSc. Thesis.

Lifeng Han, Gareth Jones, Alan Smeaton, and Paolo Bolzoni. 2021a. Chinese character decomposition for neural MT with multi-word expressions. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 336–344, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021b. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013b. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA. Association for Computational Linguistics.

Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-evaluation of Conversational Search Evaluation Metrics. *arXiv e-prints*, page arXiv:2104.13453.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Shaimaa Marzouk. 2021. An in-depth analysis of the individual impact of controlled language rules on machine translation output: a mixed-methods approach. *Machine Translation*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. *arXiv e-prints*, page arXiv:2104.07412.