In a forward direction: Analyzing distribution shifts in machine translation test sets over time

Thomas Liao [1] **Benjamin Recht** [1] **Ludwig Schmidt** [1]

Abstract

We study the effect of distribution shift between machine translation datasets by evaluating six recent English-to-German translation models on twelve years of competition test sets. We find substantial differences across years and a strong upward trend over time, even for fixed models. For the best model on the 2019 test set, the performance difference between the 2008 and 2019 test sets is three times larger than the gap between the worst and best model in our testbed. We explain this trend in terms of translationese, a well-known linguistic phenomenon. After adjusting for translationese, the performance scores across years become more comparable, but models still perform better on more recent test sets.

1. Introduction

Machine learning now often achieves impressive performance when training and test distribution agree. At the same time, current techniques still fail in unexpected and poorly understood ways when the test distribution deviates from the training data (Quionero-Candela et al., 2009; Torralba & Efros, 2011). For instance, progress on ImageNet is often cited as one of the breakthroughs in machine learning, but state-of-the-art models still see substantial performance degradation from small distribution shift (Recht et al., 2019). This raises the question if progress in other domains of machine learning is similarly brittle.

We address this question in the context of machine translation. Machine translation has seen substantial progress over the past decade and is also regarded as a key advance in machine learning. Our starting point is the widely used WMT (Workshop on Machine Translation) test sets. Since 2006, the Conference on Machine Translation (originally the WMT) has shepherded machine translation with a yearly translation competition. The organizers create a new competition test set each year, which provides a natural setting to analyze distribution shift in machine translation. Building on these datasets, the core part of our paper is a comprehensive testbed of six recent English-to-German models that we evaluate on the past twelve years of WMT test sets.

Figure 1(a) shows the main trend for each model as a function of competition year. There is a clear increase in BLEU scores *even as models are held fixed*. This plot is complementary to competitions such as ImageNet, where the test set is fixed across years and hence better performance can be attributed solely to improvements in classification models. We wish to know, therefore, to what extent improved performance scores in machine translation are due to model improvements vs. changes in datasets.

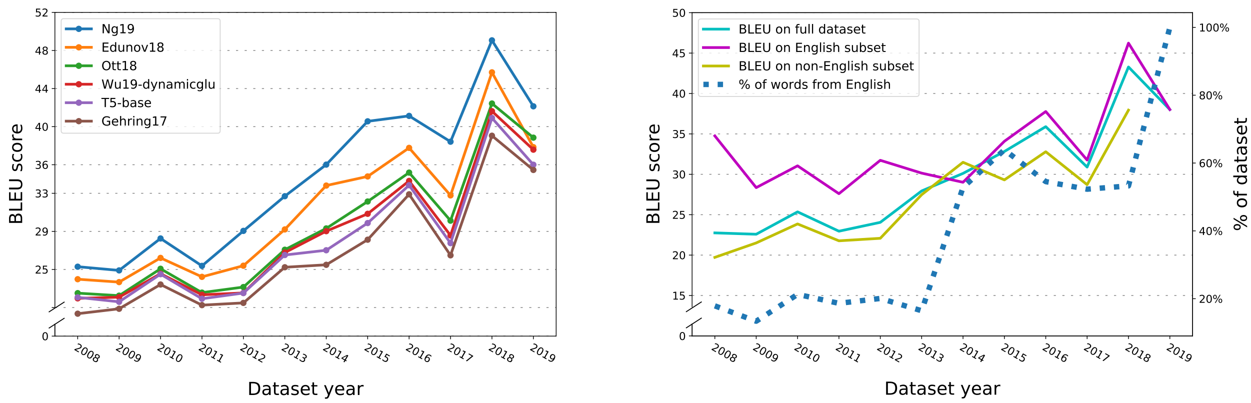
To investigate this question, we carefully dissect the WMT test sets and find that *translationese*, a known linguistic phenomenon, explains some of the trend in Figure 1(a). Translationese is text in a language X which originates in another language Y, e.g., a Czech sentence first translated into English to serve as source sentence in an English-to-German translation task, and then also translated into German to serve as target sentence. two significant changes to WMT test set construction in 2014 and 2019 aimed to control translationese. Beginning in 2014 (Bojar et al., 2014), WMT organizers stopped including most translationese sentences in the test set. And in 2019, they stopped including sentences which originated in German (Barrault et al., 2019), only including English-originating ones.

Figure 1(b) shows the aggregated performance across models after controlling for translationese. Until 2014, performance is roughly constant on the portion of each year's test set which is made of sentences originating in English. However, since these sentences comprise a small fraction of the test set, overall performance is driven by performance on the translationese component. After 2014, performance on both the English-originating subset and the non-English-originating subset rise.

Our results demonstrate that distribution shift also substantially affects machine translation models. Translationese captures some of the variations across datasets, but does not explain all performance changes. Moreover, it is unclear if

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA. Correspondence to: Thomas Liao <thomasliao@berkeley.edu>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.



(a) **Performance trends of fixed models on translation test sets**. We test six pretrained models on the past twelve years of WMT English-to-German datasets. Models achieve better BLEU on more recent test sets and perform worse on older test sets. While some models were tuned on the datasets shown here, no model was tuned on the 2019 test set.

(b) **Aggregated performance of models on test set subparts**. We bucket all model translations together to calculate aggregated performance of models on: (i) the full dataset for each year; (ii) only the sentences each year which originate in English; (iii) the remaining sentences ("translationese"). The percentage of each year's dataset originating in English is shown with a dotted line.

Figure 1: **Model performance trends on the WMT test sets from 2008 to 2019.**

the lower performance of models on translationese is due to lower data quality or because the models fail to generalize to irregular linguistic phenomena. We hope that our testbed will be a useful resource for evaluating translation models in a wide variety of contexts and making models more robust to the resulting distribution shifts.

2. Preliminaries

We restrict our attention to the English-to-German test sets (EN→DE `newstest`) for three reasons: (i) although various language pairs have been added and been removed by WMT, EN→DE has been supported since 2008, the earliest year we investigate; (ii) EN→DE is the unique language pair for which a machine translation system has been declared superior to humans in human evaluation (Toral et al., 2018; Toral, 2020; Läubli et al., 2018; 2020), an indication of the greatest actual improvement in performance; (iii) and, finally, EN→DE scores on the 2014 test set have been used as a point of comparison for a number of influential neural machine translation papers (Bahdanau et al., 2015; Vaswani et al., 2017). Further details about the test sets are elaborated in Appendix 6.1.

2.1. Models

Most of our pretrained models (Gehring et al., 2017; Edunov et al., 2018; Ott et al., 2018; Ng et al., 2019; Wu et al., 2019)) are sourced from the FAIRSEQ (Ott et al., 2019) repository and the rest (Junczys-Dowmunt et al., 2018; Lample & Conneau, 2019) are provided by the HUGGINGFACE (Wolf et al., 2019) repository. We include the WMT18 and WMT19 winners for the EN→DE direction, (Edunov et al., 2018) and (Ng et al., 2019), respectively. We provide further details on models and generation in Appendix 6.3.

2.2. BLEU scores

BLEU (Papineni et al., 2002) is the dominant metric used by the machine translation community to measure the performance of MT systems. The metric outputs a score between 0 and 100, where 100 is the maximum theoretically possible score. The state-of-the-art model on the 2019 test set achieved a BLEU of 42 and was preferred by annotators over human reference translators (Barrault et al., 2019).

All scores reported in this paper are computed using the SACREBLEU Python package[1]. More details about BLEU scores are provided in Appendix 6.2.

3. Changes in test set construction

We hypothesize that a major reason for the year-on-year score increase for fixed models is changes in test set construction — specifically, the progressive exclusion of translationese segments in 2014 and 2019 described in Section 1.

3.1. Linguistic effects of translationese

Figure 2 shows that all models perform poorly on translationese by re-grouping segments from all `newstest`

[1] `https://github.com/mjpost/SacreBLEU`

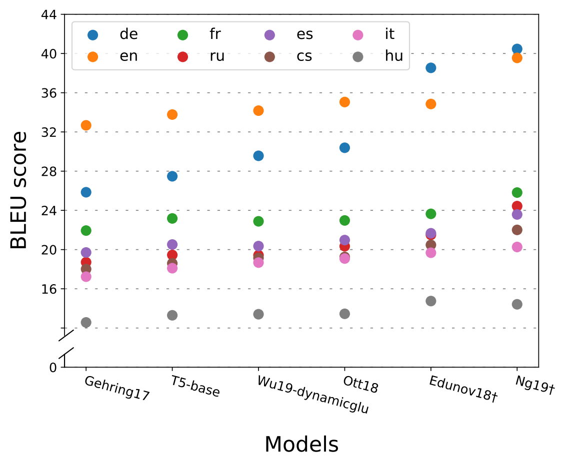
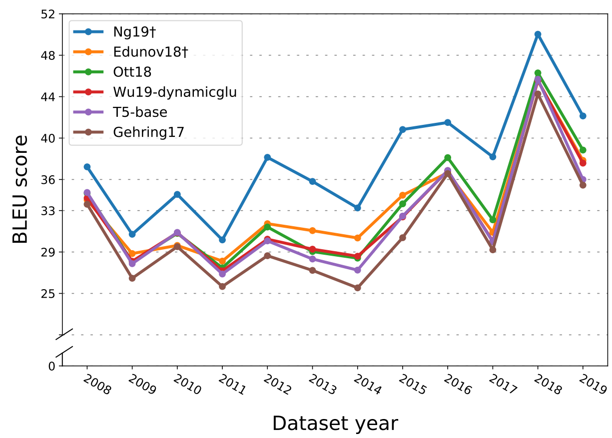


Figure 2: **Performance of fixed models on translationese**. Models are ordered on the X-axis according to performance on `newstest2019`. We re-group sentences from the past twelve years of WMT EN→DE `newstest` datasets into nine buckets based on their original language, and test models against each bucket. Models marked with '†' used back-translation during training. WMT stopped including translationese in `newstest` datasets beginning in 2019.

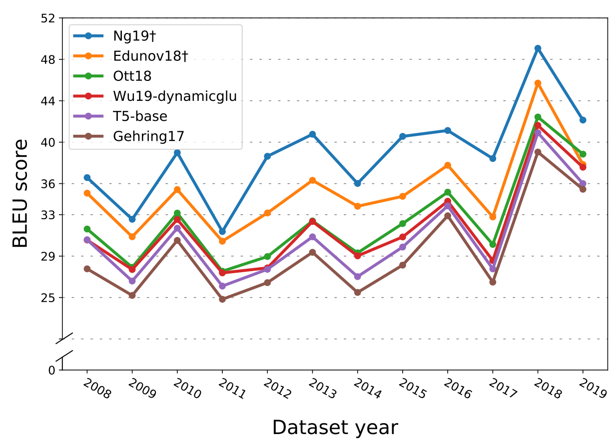
test sets based on their original language. Models are ordered on the x-axis based on their performance on `newstest2019`, with the best model on the right. The points labeled "en" represent scores for translations of segments pulled from `newstest2008,2009,...,2019` which are originally in English (i.e, the only segments which are not translationese to any degree). The points labeled "cs" represent scores for translations of segments which were originally in Czech, then translated to English for use in the EN→DE `newstests`. Not all languages were included for each year; Russian, for example, was only included in `newstest2013`.

Translationese exhibits unique artifacts from the process of translation (Baker et al., 1993), such as being simpler or being more explicit (Laviosa-Braithwaite, 1998) than the source sentence, as well as properties carried over from the source language, such as grammatical structure or over- or under-representation of particular words (Koppel & Ordan, 2011). Recent work (Graham et al., 2019) studying the effect of translationese in WMT evaluation found that translationese in WMT test sets was generally shorter than original text and that testing systems on reverse text resulted in higher human scores.

On non-German-originating translationese, models tend to perform within 5 BLEU of other, but the best model scores almost 10 BLEU higher on the originally-English segments than the worst model and nearly 15 BLEU higher on the originally-German segments. This suggests that most of the gains in English-to-German translation over time have been achieved on translating English- or German-original text, with little improvement in translating English segments that originate in a third language. All models perform significantly better on English- and German-originating segments, and worst on originally-Hungarian segments.



(a) **Model performance (BLEU) when only English-original sentences are included**.



(b) **Model performance (BLEU) when only English- or German-original sentences are included**.

Figure 3: In 2019, only English-original sentences are included, so scores are exactly the same as in Figure 1(a). Most models do worse when German-original sentences are included, except the two models trained with backtranslation (denoted with '†'). Models trained with backtranslation supplement the training data with reverse-direction sentences – sentences translated from the target language (i.e., German) into the source language (i.e., English) – so they suffer a less drastic performance drop when tested on reverse direction sentences

3.2. The effect of backtranslation

We observe that models perform best on forward direction segments, which is unsurprising when the training data consists heavily of forward-direction segments. The exception is models trained with backtranslation (denoted with '†'), which perform best on reverse direction segments.

Backtranslation (Sennrich et al., 2016a;b) is a data augmentation strategy which significantly improves model performance. Additional training sentences are generated by automatically translating sentences from monolingual corpora in the target language into the source language. Concretely for the EN→DE pair, this would entail taking sentences originally in German, translating them to English using a DE→EN translation model, then using the resulting pairs of sentences to train a EN→DE model.

We show in Figure 3 that the rising trend flattens when we remove translationese. The models marked with a '†' are trained with backtranslation, and perform better on reverse direction sentences then the other models. It is known that models trained with backtranslation are better at translating reverse direction sentences (Edunov et al., 2018) than those trained without, so it is unsurprising that they suffer a lower drop in BLEU when reverse direction sentences are added, as shown in Figure 3(b).

4. Related Work

Distribution shift in ML. Machine learning systems trained to maximize scores on particular test sets demonstrate substantial degradations in performance when tested against similar examples drawn from distributions which are similar to the original training set (Torralba & Efros, 2011). Researchers have long ascertained the performance of ML systems on test sets which remain fixed for years (e.g. the ImageNet dataset (Deng et al., 2009) for object recognition; the Penn Treebank (Marcus et al., 1993) for part-of-speech tagging). By constructing highly similar datasets drawn from similar distributions, then testing systems against them, previous work has concluded that machine learning systems are highly susceptible to minor shifts in data distributions (CIFAR-10 and ImageNet replications (Recht et al., 2019); MNIST replication (Yadav & Bottou, 2019)). Modern neural classifiers suffer a loss in accuracy equivalent to multiple years of progress. The yearly WMT test sets provide an extended chronology of natural distribution shifts.

Source-target domain mismatch. Concurrent work on backtranslation has found that the technique is less effective when there is a mismatch between the topics or domains between the source and target language (Shen et al., 2019) corpora, a problem exacerbated in low-resource language pairs. Since the EN→DE `newstest` test sets are drawn from news articles at similar points in time, we speculate that the effect of domain mismatch is substantially less than in such low-resource cases. Another independent work (Bogoychev & Sennrich, 2019) notes "subtle" domain differences in FR→EN (a high-resource pair) `newstest` test sets. They train a model to distinguish between origlang=FR and origlang=EN segments, but they admit it is ambiguous whether this model is relying on translationese artifacts or can readily distinguish between the source and target domain.

Backtranslation. (Zhang & Toral, 2019) conduct a similar experiment where they remove reverse-direction translationese sentences for `newstest2016`, `newstest2017`, `newstest2018` across a number of language pairs. They find that the best two EN→DE models in 2017 and 2018 suffer little drop in performance as judged by humans, which accords with our result that high-performing backtranslation models perform well on the reverse-direction subset on all years. In contrast to our work, they examine only models submitted for those years' competitions, whereas we score SOTA models from the full twelve years. The narrower focus on these three test sets also precludes an analysis of translationese from the third direction (e.g. origlang!=en,de for EN→DE), which we show has "weighed down" on machine translation due to greater translation difficulty.

5. Discussion and Future Work

In this work, we six state-of-the-art English-to-German translation models on twelve years of WMT translation test sets. We identify a near-doubling of BLEU scores and connect this increase to deliberate changes made by competition organizers to minimize the impact of translationese, a linguistic phenomenon caused by incorporating sentences which originate in neither the source nor target language. We recommend that researchers no longer rely exclusively on the 2014 WMT test set to compare models, as it favors models trained with a particular data augmentation, backtranslation.

Future lines of research include: creating a new set of reference translations, to control for the influence of translation quality on model performance; measuring the effect of domain shift by annotating news articles with topic or theme; soliciting human judgments of translation quality to determine whether human annotators also discern a matching increase in performance.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Baker, M., Francis, G., and Tognini-Bonelli, E. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology: In Honour of John Sinclair*. John Benjamins Publishing Company, Netherlands, 1993.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://www.aclweb.org/anthology/W19-5301.

Belz, A. and Reiter, E. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E06-1040.

Bogoychev, N. and Sennrich, R. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*, 2019.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-3302. URL https://www.aclweb.org/anthology/W14-3302.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://www.aclweb.org/anthology/W16-2301.

Callison-Burch, C., Osborne, M., and Koehn, P. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E06-1032.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, 2010. URL https://www.aclweb.org/anthology/W10-1703.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011. URL https://www.aclweb.org/anthology/W11-2103.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 2009.

Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1045. URL https://www.aclweb.org/anthology/D18-1045.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Graham, Y., Haddow, B., and Koehn, P. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*, 2019.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P18-4020.

Koppel, M. and Ordan, N. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

Lample, G. and Conneau, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Läubli, S., Sennrich, R., and Volk, M. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4792–4796, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/

v1/D18-1512. URL https://www.aclweb.org/anthology/D18-1512.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672, 2020.

Laviosa-Braithwaite, S. Universals of translation. *Routledge encyclopedia of translation studies. London: Routledge*, pp. 288–291, 1998.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 1993. URL https://www.aclweb.org/anthology/J93-2004.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-5333. URL https://www.aclweb.org/anthology/W19-5333.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. URL https://www.aclweb.org/anthology/D17-1238.

Ott, M. Personal communication, 2020.

Ott, M., Edunov, S., Grangier, D., and Auli, M. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 1–9, 2018.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018. URL https://www.aclweb.org/anthology/W18-6319.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019.

Reiter, E. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018. doi: 10.1162/coli_a_00322. URL https://www.aclweb.org/anthology/J18-3002.

Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Sennrich, R., Haddow, B., and Birch, A. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 371–376, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/W16-2323. URL https://www.aclweb.org/anthology/W16-2323.

Shen, J., Chen, P.-J., Le, M., He, J., Gu, J., Ott, M., Auli, M., and Ranzato, M. The source-target domain mismatch problem in machine translation. *arXiv preprint arXiv:1909.13151*, 2019.

Toral, A. Reassessing claims of human parity and superhuman performance in machine translation at wmt 2019. *arXiv preprint arXiv:2005.05738*, 2020.

Toral, A., Castilho, S., Hu, K., and Way, A. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 113–123, 2018.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N., and Auli, M. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.

Yadav, C. and Bottou, L. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems*, 2019.

Zhang, M. and Toral, A. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pp. 73–81, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5208. URL https://www.aclweb.org/anthology/W19-5208.

6. Appendix

6.1. Datasets

In each `newstest` instance, all sentences in the source are English sentences, and the reference sentences are rendered in German. However, WMT initially collected news articles originating in all six language of the competition: Czech, English, French, German, Hungarian, and Spanish. These articles were translated from their original language into all five other languages to create that year's test set. For example, `newstest2008` has 2051 sentences, 349 of which were originally English, 361 originally German, 416 Czech, etc. Languages were added and removed at various points. As mentioned in Section 1, only sentences originating in the source or target language (in our case, EN or DE) were used to create the test set starting in 2014 (Bojar et al., 2014). And in 2019, only sentences originating in the source language (i.e., EN) were permitted (Barrault et al., 2019).

Although translations were originally sourced from a wide mix of translators, including fluent speakers who were not professional translators, all translations have been produced since 2010 by professional translators (Callison-Burch et al., 2010). WMT organizers themselves observe that the quality of the translations has fluctuated between years (Callison-Burch et al., 2011).

Each test set is on average 2728 segments in length (a segment is a line of text, such as a sentence or a fragment); `newstest2019` is the shortest at 1997 segments, and `newstest2009` is the longest at 3027 segments. Each segment is on average 120 characters long, with a standard deviation of 68.

Language codes are as follows: Czech (CS), English (EN), French (FR), German (DE), Hungarian (HU), Russian (RU), Italian (IT), and Spanish (ES).

6.2. BLEU scores

In this paper, we always report cased detokenized BLEU (using the "v13a" tokenizer). While many papers process the reference text to split compound words, we never process the reference. SACREBLEU signatures are shown in Appendix 7.

Historically, different papers have computed BLEU inconsistently (Post, 2018), and, consequently, comparing BLEU scores between different models on the same dataset is not always kosher: changes to the scoring parameterization and processing of the reference translations can vary BLEU by as much as 1.5 points in our experience. Such irregularities mean that BLEU scores in this paper are *not* guaranteed to be commensurate with self-reported scores from prior works, and may be higher or lower than scores reported by those papers.

While substantial work (Callison-Burch et al., 2006; Novikova et al., 2017; Reiter, 2018; Bojar et al., 2016; Belz & Reiter, 2006) has cast doubt on the quality of the BLEU score as a evaluation metric (e.g. concerns over systematic divergences between human judgments of quality and BLEU scores), it has persisted as the most widely used evaluation tool in machine translation.

6.3. Translation generation

The FAIRSEQ pretrained models are provided by the authors of their respective papers and, are to the best of our knowledge, identical to the ones used for the competition, with the exception (Ott, 2020) of the model described in (Gehring et al., 2017). The XLM model (Lample & Conneau, 2019) was trained by HUGGINGFACE maintainers, not the XLM authors. Experiments were conducted on an AWS EC2 p3.16xlarge instance with a Tesla V100 GPU.

When reported, we use model-specific generation parameters, such as beam width and length penalties, but fall back on framework-specific defaults otherwise. We attempt to match text preprocessing pipelines when not provided, but exact details are rarely reported in the literature. Since even subtle differences in tokenizing punctuation marks matter for BLEU, the differences in preprocessing pipelines contributes to differences from our scores with reported scores.

7. SACREBLEU signatures

	T5-base.huggingface.en-de.19.pretrained
newstest2008	BLEU = 22.07 53.5/27.3/16.2/10.2 (BP = 0.995 ratio = 0.995 hyp_len = 47205 ref_len = 47437)
newstest2009	BLEU = 21.60 53.7/27.2/15.9/9.8 (BP = 0.989 ratio = 0.989 hyp_len = 73277 ref_len = 74087)
newstest2010	BLEU = 24.51 57.3/30.9/18.9/12.0 (BP = 0.974 ratio = 0.974 hyp_len = 59928 ref_len = 61503)
newstest2011	BLEU = 21.92 54.2/27.5/16.1/10.0 (BP = 0.990 ratio = 0.991 hyp_len = 72289 ref_len = 72981)
newstest2012	BLEU = 22.52 55.1/28.6/17.1/10.7 (BP = 0.973 ratio = 0.973 hyp_len = 70941 ref_len = 72886)
newstest2013	BLEU = 26.52 58.1/32.4/20.6/13.5 (BP = 0.986 ratio = 0.986 hyp_len = 62870 ref_len = 63737)
newstest2014	BLEU = 27.02 57.1/32.5/20.8/13.8 (BP = 1.000 ratio = 1.036 hyp_len = 64918 ref_len = 62688)
newstest2015	BLEU = 29.87 60.2/35.4/23.4/16.0 (BP = 1.000 ratio = 1.025 hyp_len = 45367 ref_len = 44260)
newstest2016	BLEU = 33.85 63.5/39.5/27.1/19.3 (BP = 1.000 ratio = 1.002 hyp_len = 62777 ref_len = 62669)
newstest2017	BLEU = 27.76 58.8/33.5/21.4/14.1 (BP = 1.000 ratio = 1.016 hyp_len = 62263 ref_len = 61287)
newstest2018	BLEU = 40.91 68.6/46.5/34.1/25.7 (BP = 1.000 ratio = 1.006 hyp_len = 64686 ref_len = 64276)
newstest2019	BLEU = 36.00 63.2/41.1/29.7/22.2 (BP = 0.995 ratio = 0.995 hyp_len = 48504 ref_len = 48746)

	Gehring17.fairseq.en-de.17.pretrained
newstest2008	BLEU = 20.35 51.8/25.3/14.6/9.0 (BP = 1.000 ratio = 1.012 hyp_len = 48023 ref_len = 47437)
newstest2009	BLEU = 20.87 53.2/26.2/15.0/9.0 (BP = 1.000 ratio = 1.002 hyp_len = 74224 ref_len = 74087)
newstest2010	BLEU = 23.42 56.9/29.9/17.8/11.1 (BP = 0.973 ratio = 0.974 hyp_len = 59890 ref_len = 61503)
newstest2011	BLEU = 21.25 53.8/26.4/15.4/9.5 (BP = 0.997 ratio = 0.997 hyp_len = 72756 ref_len = 72981)
newstest2012	BLEU = 21.49 54.7/27.3/16.0/9.9 (BP = 0.976 ratio = 0.976 hyp_len = 71132 ref_len = 72886)
newstest2013	BLEU = 25.23 57.5/31.0/19.2/12.3 (BP = 0.990 ratio = 0.990 hyp_len = 63114 ref_len = 63737)
newstest2014	BLEU = 25.49 56.4/31.0/19.3/12.5 (BP = 1.000 ratio = 1.041 hyp_len = 65261 ref_len = 62688)
newstest2015	BLEU = 28.13 59.4/33.6/21.7/14.5 (BP = 1.000 ratio = 1.026 hyp_len = 45406 ref_len = 44260)
newstest2016	BLEU = 32.89 63.5/38.6/26.1/18.3 (BP = 1.000 ratio = 1.003 hyp_len = 62854 ref_len = 62669)
newstest2017	BLEU = 26.49 58.1/32.1/20.1/13.1 (BP = 1.000 ratio = 1.027 hyp_len = 62934 ref_len = 61287)
newstest2018	BLEU = 39.05 67.7/44.7/32.2/23.9 (BP = 1.000 ratio = 1.014 hyp_len = 65163 ref_len = 64276)
newstest2019	BLEU = 35.45 63.8/40.6/28.9/21.1 (BP = 1.000 ratio = 1.004 hyp_len = 48962 ref_len = 48746)

	Ott18.fairseq.en-de.18.pretrained
newstest2008	BLEU = 22.51 54.4/28.0/16.8/10.8 (BP = 0.983 ratio = 0.983 hyp_len = 46615 ref_len = 47437)
newstest2009	BLEU = 22.25 55.3/28.3/16.8/10.4 (BP = 0.972 ratio = 0.972 hyp_len = 72042 ref_len = 74087)
newstest2010	BLEU = 25.07 59.2/32.3/19.9/12.8 (BP = 0.950 ratio = 0.951 hyp_len = 58479 ref_len = 61503)
newstest2011	BLEU = 22.57 56.0/28.6/16.9/10.6 (BP = 0.975 ratio = 0.976 hyp_len = 71196 ref_len = 72981)
newstest2012	BLEU = 23.15 57.2/29.8/17.9/11.3 (BP = 0.956 ratio = 0.957 hyp_len = 69778 ref_len = 72886)
newstest2013	BLEU = 27.07 60.1/33.5/21.4/14.0 (BP = 0.971 ratio = 0.972 hyp_len = 61923 ref_len = 63737)
newstest2014	BLEU = 29.31 60.3/35.0/22.8/15.3 (BP = 1.000 ratio = 1.008 hyp_len = 63198 ref_len = 62688)
newstest2015	BLEU = 32.14 63.3/38.0/25.5/17.7 (BP = 0.996 ratio = 0.996 hyp_len = 44081 ref_len = 44260)
newstest2016	BLEU = 35.17 66.7/41.7/28.9/20.8 (BP = 0.978 ratio = 0.978 hyp_len = 61314 ref_len = 62669)
newstest2017	BLEU = 30.13 62.0/36.1/23.6/15.8 (BP = 0.997 ratio = 0.997 hyp_len = 61100 ref_len = 61287)
newstest2018	BLEU = 42.43 71.4/48.7/35.9/27.2 (BP = 0.988 ratio = 0.988 hyp_len = 63511 ref_len = 64276)
newstest2019	BLEU = 38.84 67.2/44.5/32.6/24.5 (BP = 0.988 ratio = 0.988 hyp_len = 48164 ref_len = 48746)

	Edunov18.fairseq.en-de.18.pretrained
newstest2008	BLEU = 23.98 56.1/29.7/18.4/12.1 (BP = 0.971 ratio = 0.971 hyp_len = 46074 ref_len = 47437)
newstest2009	BLEU = 23.67 56.5/29.9/18.2/11.6 (BP = 0.970 ratio = 0.970 hyp_len = 71877 ref_len = 74087)
newstest2010	BLEU = 26.21 59.8/33.5/21.3/14.2 (BP = 0.939 ratio = 0.941 hyp_len = 57879 ref_len = 61503)
newstest2011	BLEU = 24.22 57.5/30.5/18.6/12.0 (BP = 0.967 ratio = 0.968 hyp_len = 70622 ref_len = 72981)
newstest2012	BLEU = 25.40 59.0/32.3/20.3/13.3 (BP = 0.947 ratio = 0.949 hyp_len = 69135 ref_len = 72886)
newstest2013	BLEU = 29.21 61.5/35.8/23.6/16.1 (BP = 0.965 ratio = 0.966 hyp_len = 61561 ref_len = 63737)
newstest2014	BLEU = 33.81 63.5/39.6/27.2/19.2 (BP = 0.998 ratio = 0.998 hyp_len = 62554 ref_len = 62688)
newstest2015	BLEU = 34.77 65.3/40.8/28.2/20.1 (BP = 0.992 ratio = 0.992 hyp_len = 43900 ref_len = 44260)
newstest2016	BLEU = 37.77 69.0/45.2/32.1/23.6 (BP = 0.964 ratio = 0.964 hyp_len = 60440 ref_len = 62669)
newstest2017	BLEU = 32.78 64.9/39.3/26.6/18.6 (BP = 0.979 ratio = 0.979 hyp_len = 59991 ref_len = 61287)
newstest2018	BLEU = 45.70 73.9/52.6/40.0/31.2 (BP = 0.974 ratio = 0.974 hyp_len = 62604 ref_len = 64276)
newstest2019	BLEU = 37.85 66.6/43.9/32.0/24.2 (BP = 0.975 ratio = 0.976 hyp_len = 47552 ref_len = 48746)

	Ng19.fairseq.en-de.19.pretrained
newstest2008	BLEU = 25.29 56.6/30.7/19.2/12.7 (BP = 0.991 ratio = 0.991 hyp_len = 47002 ref_len = 47437)
newstest2009	BLEU = 24.90 56.9/30.6/18.9/12.3 (BP = 0.987 ratio = 0.987 hyp_len = 73122 ref_len = 74087)
newstest2010	BLEU = 28.26 61.5/35.3/22.9/15.5 (BP = 0.953 ratio = 0.954 hyp_len = 58691 ref_len = 61503)
newstest2011	BLEU = 25.38 58.0/31.3/19.4/12.7 (BP = 0.982 ratio = 0.982 hyp_len = 71656 ref_len = 72981)
newstest2012	BLEU = 29.05 61.5/35.8/23.5/16.2 (BP = 0.960 ratio = 0.961 hyp_len = 70053 ref_len = 72886)
newstest2013	BLEU = 32.70 63.8/39.0/26.7/18.9 (BP = 0.976 ratio = 0.976 hyp_len = 62229 ref_len = 63737)
newstest2014	BLEU = 36.01 65.0/41.7/29.3/21.2 (BP = 1.000 ratio = 1.011 hyp_len = 63403 ref_len = 62688)
newstest2015	BLEU = 40.56 68.9/46.2/33.7/25.3 (BP = 1.000 ratio = 1.000 hyp_len = 44244 ref_len = 44260)
newstest2016	BLEU = 41.13 70.9/47.8/34.8/26.1 (BP = 0.982 ratio = 0.982 hyp_len = 61539 ref_len = 62669)
newstest2017	BLEU = 38.42 68.3/44.4/31.7/23.3 (BP = 0.993 ratio = 0.993 hyp_len = 60830 ref_len = 61287)
newstest2018	BLEU = 49.07 75.6/55.2/42.8/33.9 (BP = 0.989 ratio = 0.989 hyp_len = 63559 ref_len = 64276)
newstest2019	BLEU = 42.14 69.7/47.7/35.6/27.3 (BP = 0.994 ratio = 0.994 hyp_len = 48461 ref_len = 48746)

	Wu19-dynamicglu.fairseq.en-de.16.pretrained
newstest2008	BLEU = 21.96 54.1/27.5/16.4/10.4 (BP = 0.980 ratio = 0.980 hyp_len = 46480 ref_len = 47437)
newstest2009	BLEU = 22.09 55.3/28.5/16.9/10.4 (BP = 0.963 ratio = 0.964 hyp_len = 71416 ref_len = 74087)
newstest2010	BLEU = 24.56 58.6/31.7/19.4/12.5 (BP = 0.947 ratio = 0.948 hyp_len = 58317 ref_len = 61503)
newstest2011	BLEU = 22.34 55.4/28.4/16.8/10.5 (BP = 0.972 ratio = 0.973 hyp_len = 70991 ref_len = 72981)
newstest2012	BLEU = 22.53 56.3/29.3/17.5/11.0 (BP = 0.950 ratio = 0.951 hyp_len = 69322 ref_len = 72886)
newstest2013	BLEU = 26.77 59.5/33.2/21.2/14.0 (BP = 0.967 ratio = 0.968 hyp_len = 61697 ref_len = 63737)
newstest2014	BLEU = 29.02 59.5/34.7/22.6/15.2 (BP = 1.000 ratio = 1.014 hyp_len = 63546 ref_len = 62688)
newstest2015	BLEU = 30.85 62.1/36.7/24.3/16.7 (BP = 0.995 ratio = 0.995 hyp_len = 44027 ref_len = 44260)
newstest2016	BLEU = 34.30 65.5/41.0/28.3/20.2 (BP = 0.974 ratio = 0.974 hyp_len = 61054 ref_len = 62669)
newstest2017	BLEU = 28.57 60.4/34.5/22.2/14.7 (BP = 0.995 ratio = 0.995 hyp_len = 61002 ref_len = 61287)
newstest2018	BLEU = 41.62 70.0/47.8/35.3/26.7 (BP = 0.987 ratio = 0.987 hyp_len = 63444 ref_len = 64276)
newstest2019	BLEU = 37.58 65.5/43.2/31.5/23.7 (BP = 0.987 ratio = 0.987 hyp_len = 48096 ref_len = 48746)

	Aggregated performance of all models on full dataset
newstest2008	BLEU = 22.75 54.4/28.1/16.9/10.8 (BP = 0.989 ratio = 0.989 hyp_len = 281399 ref_len = 284622)
newstest2009	BLEU = 22.58 55.2/28.4/17.0/10.6 (BP = 0.981 ratio = 0.981 hyp_len = 435958 ref_len = 444522)
newstest2010	BLEU = 25.35 58.8/32.2/20.0/13.0 (BP = 0.956 ratio = 0.957 hyp_len = 353184 ref_len = 369018)
newstest2011	BLEU = 22.96 55.8/28.8/17.2/10.9 (BP = 0.981 ratio = 0.981 hyp_len = 429510 ref_len = 437886)
newstest2012	BLEU = 24.05 57.3/30.5/18.7/12.1 (BP = 0.960 ratio = 0.961 hyp_len = 420361 ref_len = 437316)
newstest2013	BLEU = 27.94 60.0/34.1/22.1/14.8 (BP = 0.976 ratio = 0.976 hyp_len = 373394 ref_len = 382422)
newstest2014	BLEU = 30.11 60.3/35.7/23.6/16.2 (BP = 1.000 ratio = 1.018 hyp_len = 382880 ref_len = 376128)
newstest2015	BLEU = 32.82 63.2/38.4/26.1/18.3 (BP = 1.000 ratio = 1.006 hyp_len = 267025 ref_len = 265560)
newstest2016	BLEU = 35.90 66.5/42.3/29.5/21.3 (BP = 0.984 ratio = 0.984 hyp_len = 369978 ref_len = 376014)
newstest2017	BLEU = 30.89 62.1/36.6/24.2/16.6 (BP = 1.000 ratio = 1.001 hyp_len = 368120 ref_len = 367722)
newstest2018	BLEU = 43.28 71.2/49.2/36.7/28.1 (BP = 0.993 ratio = 0.993 hyp_len = 382967 ref_len = 385656)
newstest2019	BLEU = 38.01 66.0/43.5/31.7/23.8 (BP = 0.991 ratio = 0.991 hyp_len = 289739 ref_len = 292476)

	T5-base.huggingface.en-de.19.pretrained
origlang=en newstest2008	BLEU = 34.75 66.6/42.0/28.5/19.9 (BP = 0.979 ratio = 0.979 hyp_len = 8812 ref_len = 9002)
origlang=en newstest2009	BLEU = 27.87 60.9/34.9/22.2/14.6 (BP = 0.967 ratio = 0.968 hyp_len = 10513 ref_len = 10866)
origlang=en newstest2010	BLEU = 30.89 61.6/37.5/24.9/17.1 (BP = 0.981 ratio = 0.981 hyp_len = 13294 ref_len = 13554)
origlang=en newstest2011	BLEU = 26.87 55.9/31.7/20.7/14.2 (BP = 1.000 ratio = 1.007 hyp_len = 14220 ref_len = 14123)
origlang=en newstest2012	BLEU = 30.07 61.5/36.8/24.5/16.8 (BP = 0.968 ratio = 0.969 hyp_len = 14794 ref_len = 15268)
origlang=en newstest2013	BLEU = 28.32 59.3/34.0/22.4/15.4 (BP = 0.980 ratio = 0.981 hyp_len = 10705 ref_len = 10916)
origlang=en newstest2014	BLEU = 27.25 57.2/32.7/21.1/14.0 (BP = 1.000 ratio = 1.010 hyp_len = 36096 ref_len = 35745)
origlang=en newstest2015	BLEU = 32.47 63.1/38.3/25.8/17.9 (BP = 1.000 ratio = 1.004 hyp_len = 30334 ref_len = 30207)
origlang=en newstest2016	BLEU = 36.87 66.4/43.4/30.9/22.7 (BP = 0.978 ratio = 0.978 hyp_len = 35858 ref_len = 36655)
origlang=en newstest2017	BLEU = 29.98 62.2/36.8/24.0/16.1 (BP = 0.978 ratio = 0.979 hyp_len = 33574 ref_len = 34310)
origlang=en newstest2018	BLEU = 45.68 73.0/52.5/40.1/31.2 (BP = 0.976 ratio = 0.977 hyp_len = 36362 ref_len = 37232)
origlang=en newstest2019	BLEU = 36.00 63.2/41.1/29.7/22.2 (BP = 0.995 ratio = 0.995 hyp_len = 48504 ref_len = 48746)

	Gehring17.fairseq.en-de.17.pretrained
origlang=en newstest2008	BLEU = 33.61 65.6/40.4/27.1/18.6 (BP = 0.990 ratio = 0.990 hyp_len = 8911 ref_len = 9002)
origlang=en newstest2009	BLEU = 26.46 60.3/33.5/20.6/13.2 (BP = 0.972 ratio = 0.973 hyp_len = 10569 ref_len = 10866)
origlang=en newstest2010	BLEU = 29.51 62.0/36.5/23.5/15.7 (BP = 0.977 ratio = 0.977 hyp_len = 13241 ref_len = 13554)
origlang=en newstest2011	BLEU = 25.67 56.3/30.5/19.4/13.1 (BP = 1.000 ratio = 1.017 hyp_len = 14359 ref_len = 14123)
origlang=en newstest2012	BLEU = 28.64 61.4/35.1/22.5/14.9 (BP = 0.982 ratio = 0.982 hyp_len = 14997 ref_len = 15268)
origlang=en newstest2013	BLEU = 27.22 58.7/32.8/21.2/14.3 (BP = 0.985 ratio = 0.985 hyp_len = 10754 ref_len = 10916)
origlang=en newstest2014	BLEU = 25.54 56.8/31.2/19.4/12.4 (BP = 1.000 ratio = 1.011 hyp_len = 36130 ref_len = 35745)
origlang=en newstest2015	BLEU = 30.38 62.4/36.2/23.8/16.0 (BP = 0.997 ratio = 0.997 hyp_len = 30127 ref_len = 30207)
origlang=en newstest2016	BLEU = 36.57 67.6/43.4/30.5/22.1 (BP = 0.976 ratio = 0.977 hyp_len = 35801 ref_len = 36655)
origlang=en newstest2017	BLEU = 29.21 62.2/35.7/22.8/15.2 (BP = 0.985 ratio = 0.985 hyp_len = 33801 ref_len = 34310)
origlang=en newstest2018	BLEU = 44.26 72.8/51.0/38.2/29.3 (BP = 0.980 ratio = 0.980 hyp_len = 36501 ref_len = 37232)
origlang=en newstest2019	BLEU = 35.45 63.8/40.6/28.9/21.1 (BP = 1.000 ratio = 1.004 hyp_len = 48962 ref_len = 48746)

	Ott18.fairseq.en-de.18.pretrained
origlang=en newstest2008	BLEU = 34.67 67.7/42.4/28.7/20.0 (BP = 0.968 ratio = 0.968 hyp_len = 8716 ref_len = 9002)
origlang=en newstest2009	BLEU = 28.06 62.9/36.1/22.6/14.8 (BP = 0.951 ratio = 0.952 hyp_len = 10345 ref_len = 10866)
origlang=en newstest2010	BLEU = 30.80 64.2/38.4/25.1/17.2 (BP = 0.959 ratio = 0.959 hyp_len = 13004 ref_len = 13554)
origlang=en newstest2011	BLEU = 27.47 58.0/32.6/21.1/14.3 (BP = 1.000 ratio = 1.003 hyp_len = 14161 ref_len = 14123)
origlang=en newstest2012	BLEU = 31.41 64.3/38.6/25.4/17.4 (BP = 0.971 ratio = 0.971 hyp_len = 14832 ref_len = 15268)
origlang=en newstest2013	BLEU = 29.04 61.6/35.3/23.2/15.8 (BP = 0.972 ratio = 0.973 hyp_len = 10616 ref_len = 10916)
origlang=en newstest2014	BLEU = 28.41 59.8/34.3/22.2/14.7 (BP = 0.993 ratio = 0.993 hyp_len = 35499 ref_len = 35745)
origlang=en newstest2015	BLEU = 33.65 65.8/40.5/27.4/19.1 (BP = 0.979 ratio = 0.979 hyp_len = 29585 ref_len = 30207)
origlang=en newstest2016	BLEU = 38.11 69.7/45.5/32.5/23.9 (BP = 0.962 ratio = 0.962 hyp_len = 35276 ref_len = 36655)
origlang=en newstest2017	BLEU = 32.10 65.3/39.5/26.1/17.8 (BP = 0.971 ratio = 0.972 hyp_len = 33333 ref_len = 34310)
origlang=en newstest2018	BLEU = 46.29 75.4/53.8/41.0/31.8 (BP = 0.965 ratio = 0.966 hyp_len = 35968 ref_len = 37232)
origlang=en newstest2019	BLEU = 38.84 67.2/44.5/32.6/24.5 (BP = 0.988 ratio = 0.988 hyp_len = 48164 ref_len = 48746)

	Edunov18.fairseq.en-de.18.pretrained
origlang=en newstest2008	BLEU = 34.03 66.3/41.3/28.2/19.9 (BP = 0.966 ratio = 0.967 hyp_len = 8705 ref_len = 9002)
origlang=en newstest2009	BLEU = 28.85 63.4/36.7/23.7/15.9 (BP = 0.944 ratio = 0.945 hyp_len = 10272 ref_len = 10866)
origlang=en newstest2010	BLEU = 29.63 61.0/36.5/24.3/16.8 (BP = 0.959 ratio = 0.959 hyp_len = 13003 ref_len = 13554)
origlang=en newstest2011	BLEU = 28.12 58.3/33.1/21.7/14.9 (BP = 1.000 ratio = 1.002 hyp_len = 14152 ref_len = 14123)
origlang=en newstest2012	BLEU = 31.73 64.1/38.7/26.3/18.3 (BP = 0.961 ratio = 0.961 hyp_len = 14678 ref_len = 15268)
origlang=en newstest2013	BLEU = 31.06 63.0/37.9/25.3/17.7 (BP = 0.965 ratio = 0.966 hyp_len = 10544 ref_len = 10916)
origlang=en newstest2014	BLEU = 30.35 61.1/36.5/24.4/16.6 (BP = 0.984 ratio = 0.984 hyp_len = 35185 ref_len = 35745)
origlang=en newstest2015	BLEU = 34.48 66.4/41.3/28.4/20.1 (BP = 0.976 ratio = 0.976 hyp_len = 29493 ref_len = 30207)
origlang=en newstest2016	BLEU = 36.66 68.7/44.6/31.5/23.0 (BP = 0.950 ratio = 0.951 hyp_len = 34857 ref_len = 36655)
origlang=en newstest2017	BLEU = 30.90 65.2/38.5/25.3/17.2 (BP = 0.955 ratio = 0.956 hyp_len = 32809 ref_len = 34310)
origlang=en newstest2018	BLEU = 45.53 74.7/53.4/40.9/32.1 (BP = 0.952 ratio = 0.953 hyp_len = 35484 ref_len = 37232)
origlang=en newstest2019	BLEU = 37.85 66.6/43.9/32.0/24.2 (BP = 0.975 ratio = 0.976 hyp_len = 47552 ref_len = 48746)

	Ng19.fairseq.en-de.19.pretrained
origlang=en newstest2008	BLEU = 37.22 69.1/44.5/31.1/22.1 (BP = 0.976 ratio = 0.976 hyp_len = 8786 ref_len = 9002)
origlang=en newstest2009	BLEU = 30.71 64.5/38.1/24.9/16.9 (BP = 0.964 ratio = 0.964 hyp_len = 10480 ref_len = 10866)
origlang=en newstest2010	BLEU = 34.56 67.2/42.0/28.5/20.1 (BP = 0.969 ratio = 0.969 hyp_len = 13134 ref_len = 13554)
origlang=en newstest2011	BLEU = 30.17 60.1/35.2/23.6/16.6 (BP = 1.000 ratio = 1.009 hyp_len = 14247 ref_len = 14123)
origlang=en newstest2012	BLEU = 38.14 68.8/44.9/32.0/23.5 (BP = 0.977 ratio = 0.978 hyp_len = 14927 ref_len = 15268)
origlang=en newstest2013	BLEU = 35.81 65.8/42.0/29.8/21.8 (BP = 0.979 ratio = 0.979 hyp_len = 10686 ref_len = 10916)
origlang=en newstest2014	BLEU = 33.25 63.1/39.0/26.7/18.6 (BP = 1.000 ratio = 1.001 hyp_len = 35775 ref_len = 35745)
origlang=en newstest2015	BLEU = 40.83 70.3/47.3/34.4/25.7 (BP = 0.986 ratio = 0.986 hyp_len = 29774 ref_len = 30207)
origlang=en newstest2016	BLEU = 41.51 71.7/48.8/35.5/26.6 (BP = 0.974 ratio = 0.974 hyp_len = 35704 ref_len = 36655)
origlang=en newstest2017	BLEU = 38.19 69.8/45.3/32.0/23.2 (BP = 0.976 ratio = 0.977 hyp_len = 33505 ref_len = 34310)
origlang=en newstest2018	BLEU = 50.03 77.1/56.9/44.7/35.7 (BP = 0.973 ratio = 0.973 hyp_len = 36242 ref_len = 37232)
origlang=en newstest2019	BLEU = 42.14 69.7/47.7/35.6/27.3 (BP = 0.994 ratio = 0.994 hyp_len = 48461 ref_len = 48746)

	Wu19-dynamicglu.fairseq.en-de.16.pretrained
origlang=en newstest2008	BLEU = 34.21 66.6/41.8/28.2/19.6 (BP = 0.971 ratio = 0.972 hyp_len = 8747 ref_len = 9002)
origlang=en newstest2009	BLEU = 28.10 62.2/35.9/22.8/15.1 (BP = 0.949 ratio = 0.950 hyp_len = 10326 ref_len = 10866)
origlang=en newstest2010	BLEU = 30.85 63.6/38.4/25.5/17.8 (BP = 0.951 ratio = 0.952 hyp_len = 12907 ref_len = 13554)
origlang=en newstest2011	BLEU = 27.18 56.8/32.1/21.0/14.5 (BP = 0.996 ratio = 0.996 hyp_len = 14068 ref_len = 14123)
origlang=en newstest2012	BLEU = 30.24 62.4/37.4/24.7/17.0 (BP = 0.961 ratio = 0.962 hyp_len = 14687 ref_len = 15268)
origlang=en newstest2013	BLEU = 29.27 60.8/35.5/23.4/15.9 (BP = 0.978 ratio = 0.979 hyp_len = 10682 ref_len = 10916)
origlang=en newstest2014	BLEU = 28.59 59.4/34.4/22.4/15.0 (BP = 0.994 ratio = 0.994 hyp_len = 35518 ref_len = 35745)
origlang=en newstest2015	BLEU = 32.39 64.8/39.2/26.2/18.1 (BP = 0.978 ratio = 0.978 hyp_len = 29538 ref_len = 30207)
origlang=en newstest2016	BLEU = 36.86 68.1/44.4/31.6/23.2 (BP = 0.956 ratio = 0.957 hyp_len = 35062 ref_len = 36655)
origlang=en newstest2017	BLEU = 30.04 63.3/37.3/24.3/16.4 (BP = 0.964 ratio = 0.965 hyp_len = 33096 ref_len = 34310)
origlang=en newstest2018	BLEU = 45.56 74.1/53.2/40.6/31.5 (BP = 0.962 ratio = 0.962 hyp_len = 35831 ref_len = 37232)
origlang=en newstest2019	BLEU = 37.58 65.5/43.2/31.5/23.7 (BP = 0.987 ratio = 0.987 hyp_len = 48096 ref_len = 48746)

	T5-base.huggingface.en-de.19.pretrained
origlang={en,de} newstest2008	BLEU = 30.58 62.1/36.8/24.0/16.2 (BP = 0.995 ratio = 0.995 hyp_len = 16042 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 26.59 57.7/32.3/20.3/13.2 (BP = 1.000 ratio = 1.003 hyp_len = 18967 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 31.69 61.9/37.8/25.2/17.2 (BP = 1.000 ratio = 1.006 hyp_len = 23297 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 26.10 56.2/31.2/19.8/13.3 (BP = 1.000 ratio = 1.021 hyp_len = 25816 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 27.72 58.7/33.5/21.5/14.3 (BP = 0.994 ratio = 0.994 hyp_len = 26193 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 30.86 61.1/36.5/24.3/16.7 (BP = 1.000 ratio = 1.013 hyp_len = 19781 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 27.02 57.1/32.5/20.8/13.8 (BP = 1.000 ratio = 1.036 hyp_len = 64918 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 29.87 60.2/35.4/23.4/16.0 (BP = 1.000 ratio = 1.025 hyp_len = 45367 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 33.85 63.5/39.5/27.1/19.3 (BP = 1.000 ratio = 1.002 hyp_len = 62777 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 27.76 58.8/33.5/21.4/14.1 (BP = 1.000 ratio = 1.016 hyp_len = 62263 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 40.91 68.6/46.5/34.1/25.7 (BP = 1.000 ratio = 1.006 hyp_len = 64686 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 36.00 63.2/41.1/29.7/22.2 (BP = 0.995 ratio = 0.995 hyp_len = 48504 ref_len = 48746)

	Gehring17.fairseq.en-de.17.pretrained
origlang={en,de} newstest2008	BLEU = 27.77 59.0/33.5/21.3/14.1 (BP = 1.000 ratio = 1.028 hyp_len = 16574 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 25.20 57.1/31.0/19.0/12.0 (BP = 1.000 ratio = 1.015 hyp_len = 19193 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 30.51 62.2/36.8/23.9/16.0 (BP = 0.998 ratio = 0.998 hyp_len = 23100 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 24.83 56.1/29.8/18.6/12.2 (BP = 1.000 ratio = 1.030 hyp_len = 26040 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 26.44 58.5/32.1/20.1/13.1 (BP = 0.997 ratio = 0.997 hyp_len = 26277 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 29.36 60.4/35.0/22.9/15.4 (BP = 1.000 ratio = 1.019 hyp_len = 19899 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 25.49 56.4/31.0/19.3/12.5 (BP = 1.000 ratio = 1.041 hyp_len = 65261 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 28.13 59.4/33.6/21.7/14.5 (BP = 1.000 ratio = 1.026 hyp_len = 45406 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 32.89 63.5/38.6/26.1/18.3 (BP = 1.000 ratio = 1.003 hyp_len = 62854 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 26.49 58.1/32.1/20.1/13.1 (BP = 1.000 ratio = 1.027 hyp_len = 62934 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 39.05 67.7/44.7/32.2/23.9 (BP = 1.000 ratio = 1.014 hyp_len = 65163 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 35.45 63.8/40.6/28.9/21.1 (BP = 1.000 ratio = 1.004 hyp_len = 48962 ref_len = 48746)

	Ott18.fairseq.en-de.18.pretrained
origlang={en,de} newstest2008	BLEU = 31.61 63.6/38.2/25.1/17.1 (BP = 0.989 ratio = 0.989 hyp_len = 15935 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 27.92 60.5/34.4/21.9/14.3 (BP = 0.982 ratio = 0.982 hyp_len = 18577 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 33.15 65.3/40.2/27.0/18.7 (BP = 0.977 ratio = 0.977 hyp_len = 22615 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 27.54 58.9/33.0/21.0/14.1 (BP = 1.000 ratio = 1.009 hyp_len = 25495 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 28.95 61.6/35.4/22.9/15.4 (BP = 0.978 ratio = 0.979 hyp_len = 25783 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 32.39 63.8/38.3/25.7/17.7 (BP = 0.997 ratio = 0.997 hyp_len = 19465 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 29.31 60.3/35.0/22.8/15.3 (BP = 1.000 ratio = 1.008 hyp_len = 63198 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 32.14 63.3/38.0/25.5/17.7 (BP = 0.996 ratio = 0.996 hyp_len = 44081 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 35.17 66.7/41.7/28.9/20.8 (BP = 0.978 ratio = 0.978 hyp_len = 61314 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 30.13 62.0/36.1/23.6/15.8 (BP = 0.997 ratio = 0.997 hyp_len = 61100 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 42.43 71.4/48.7/35.9/27.2 (BP = 0.988 ratio = 0.988 hyp_len = 63511 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 38.84 67.2/44.5/32.6/24.5 (BP = 0.988 ratio = 0.988 hyp_len = 48164 ref_len = 48746)

	Edunov18.fairseq.en-de.18.pretrained
origlang={en,de} newstest2008	BLEU = 35.08 66.1/41.5/28.8/20.8 (BP = 0.980 ratio = 0.981 hyp_len = 15804 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 30.87 63.3/37.7/25.0/17.1 (BP = 0.971 ratio = 0.972 hyp_len = 18376 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 35.42 65.7/42.2/29.7/21.4 (BP = 0.973 ratio = 0.973 hyp_len = 22523 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 30.44 60.8/35.7/23.8/16.6 (BP = 1.000 ratio = 1.004 hyp_len = 25366 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 33.16 64.8/39.8/27.4/19.3 (BP = 0.970 ratio = 0.970 hyp_len = 25567 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 36.32 66.3/42.4/29.8/21.7 (BP = 0.990 ratio = 0.990 hyp_len = 19320 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 33.81 63.5/39.6/27.2/19.2 (BP = 0.998 ratio = 0.998 hyp_len = 62554 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 34.77 65.3/40.8/28.2/20.1 (BP = 0.992 ratio = 0.992 hyp_len = 43900 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 37.77 69.0/45.2/32.1/23.6 (BP = 0.964 ratio = 0.964 hyp_len = 60440 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 32.78 64.9/39.3/26.6/18.6 (BP = 0.979 ratio = 0.979 hyp_len = 59991 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 45.70 73.9/52.6/40.0/31.2 (BP = 0.974 ratio = 0.974 hyp_len = 62604 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 37.85 66.6/43.9/32.0/24.2 (BP = 0.975 ratio = 0.976 hyp_len = 47552 ref_len = 48746)

	Ng19.fairseq.en-de.19.pretrained
origlang={en,de} newstest2008	BLEU = 36.58 67.2/42.8/29.8/21.4 (BP = 0.995 ratio = 0.995 hyp_len = 16033 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 32.55 64.2/38.9/26.1/18.0 (BP = 0.989 ratio = 0.989 hyp_len = 18699 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 38.99 69.3/45.8/32.6/24.0 (BP = 0.982 ratio = 0.982 hyp_len = 22732 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 31.37 61.6/36.6/24.7/17.4 (BP = 1.000 ratio = 1.014 hyp_len = 25626 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 38.64 68.5/45.1/32.5/24.1 (BP = 0.980 ratio = 0.980 hyp_len = 25831 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 40.76 69.0/46.4/33.9/25.4 (BP = 1.000 ratio = 1.000 hyp_len = 19518 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 36.01 65.0/41.7/29.3/21.2 (BP = 1.000 ratio = 1.011 hyp_len = 63403 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 40.56 68.9/46.2/33.7/25.3 (BP = 1.000 ratio = 1.000 hyp_len = 44244 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 41.13 70.9/47.8/34.8/26.1 (BP = 0.982 ratio = 0.982 hyp_len = 61539 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 38.42 68.3/44.4/31.7/23.3 (BP = 0.993 ratio = 0.993 hyp_len = 60830 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 49.07 75.6/55.2/42.8/33.9 (BP = 0.989 ratio = 0.989 hyp_len = 63559 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 42.14 69.7/47.7/35.6/27.3 (BP = 0.994 ratio = 0.994 hyp_len = 48461 ref_len = 48746)

	Wu19-dynamicglu.fairseq.en-de.16.pretrained
origlang={en,de} newstest2008	BLEU = 30.55 62.4/37.0/24.1/16.3 (BP = 0.990 ratio = 0.990 hyp_len = 15962 ref_len = 16116)
origlang={en,de} newstest2009	BLEU = 27.70 60.1/34.2/21.7/14.3 (BP = 0.980 ratio = 0.981 hyp_len = 18547 ref_len = 18914)
origlang={en,de} newstest2010	BLEU = 32.54 64.5/39.6/26.5/18.5 (BP = 0.973 ratio = 0.973 hyp_len = 22534 ref_len = 23151)
origlang={en,de} newstest2011	BLEU = 27.38 57.8/32.6/21.0/14.2 (BP = 1.000 ratio = 1.000 hyp_len = 25284 ref_len = 25273)
origlang={en,de} newstest2012	BLEU = 27.86 60.2/34.6/22.3/15.0 (BP = 0.964 ratio = 0.964 hyp_len = 25409 ref_len = 26348)
origlang={en,de} newstest2013	BLEU = 32.34 62.8/38.2/25.7/17.8 (BP = 1.000 ratio = 1.004 hyp_len = 19595 ref_len = 19519)
origlang={en,de} newstest2014	BLEU = 29.02 59.5/34.7/22.6/15.2 (BP = 1.000 ratio = 1.014 hyp_len = 63546 ref_len = 62688)
origlang={en,de} newstest2015	BLEU = 30.85 62.1/36.7/24.3/16.7 (BP = 0.995 ratio = 0.995 hyp_len = 44027 ref_len = 44260)
origlang={en,de} newstest2016	BLEU = 34.30 65.5/41.0/28.3/20.2 (BP = 0.974 ratio = 0.974 hyp_len = 61054 ref_len = 62669)
origlang={en,de} newstest2017	BLEU = 28.57 60.4/34.5/22.2/14.7 (BP = 0.995 ratio = 0.995 hyp_len = 61002 ref_len = 61287)
origlang={en,de} newstest2018	BLEU = 41.62 70.0/47.8/35.3/26.7 (BP = 0.987 ratio = 0.987 hyp_len = 63444 ref_len = 64276)
origlang={en,de} newstest2019	BLEU = 37.58 65.5/43.2/31.5/23.7 (BP = 0.987 ratio = 0.987 hyp_len = 48096 ref_len = 48746)

	T5-base.huggingface.en-de.19.pretrained
de	BLEU = 27.48 58.1/33.1/21.1/14.1 (BP = 1.000 ratio = 1.051 hyp_len = 185545 ref_len = 176623)
en	BLEU = 33.77 63.3/39.7/27.5/19.7 (BP = 0.988 ratio = 0.988 hyp_len = 293066 ref_len = 296624)
fr	BLEU = 23.18 57.0/30.1/18.0/11.3 (BP = 0.955 ratio = 0.956 hyp_len = 68613 ref_len = 71785)
ru	BLEU = 19.45 51.3/24.6/14.1/8.2 (BP = 0.993 ratio = 0.993 hyp_len = 9266 ref_len = 9329)
es	BLEU = 20.52 54.0/26.7/15.3/9.2 (BP = 0.966 ratio = 0.967 hyp_len = 79113 ref_len = 81826)
cz	BLEU = 20.02 53.2/25.9/14.5/8.4 (BP = 0.988 ratio = 0.988 hyp_len = 42313 ref_len = 42812)
cs	BLEU = 17.17 50.2/23.2/12.6/7.2 (BP = 0.953 ratio = 0.954 hyp_len = 24921 ref_len = 26113)
it	BLEU = 18.10 50.5/23.4/12.7/7.1 (BP = 1.000 ratio = 1.021 hyp_len = 10447 ref_len = 10229)
hu	BLEU = 13.30 43.9/17.5/8.8/4.7 (BP = 1.000 ratio = 1.025 hyp_len = 21741 ref_len = 21216)

	Gehring17.fairseq.en-de.17.pretrained
de	BLEU = 25.85 56.7/31.3/19.6/12.8 (BP = 1.000 ratio = 1.062 hyp_len = 187510 ref_len = 176623)
en	BLEU = 32.68 63.4/38.6/26.2/18.4 (BP = 0.992 ratio = 0.992 hyp_len = 294153 ref_len = 296624)
fr	BLEU = 21.94 56.2/28.8/16.7/10.1 (BP = 0.959 ratio = 0.960 hyp_len = 68879 ref_len = 71785)
ru	BLEU = 18.71 50.7/23.6/13.3/7.8 (BP = 0.998 ratio = 0.998 hyp_len = 9306 ref_len = 9329)
es	BLEU = 19.71 53.3/25.4/14.4/8.6 (BP = 0.974 ratio = 0.974 hyp_len = 79692 ref_len = 81826)
cz	BLEU = 19.64 52.9/25.3/14.0/8.2 (BP = 0.991 ratio = 0.991 hyp_len = 42447 ref_len = 42812)
cs	BLEU = 16.38 50.0/22.1/11.7/6.6 (BP = 0.957 ratio = 0.958 hyp_len = 25020 ref_len = 26113)
it	BLEU = 17.24 49.5/22.2/11.9/6.7 (BP = 1.000 ratio = 1.029 hyp_len = 10526 ref_len = 10229)
hu	BLEU = 12.57 43.1/16.8/8.1/4.2 (BP = 1.000 ratio = 1.046 hyp_len = 22186 ref_len = 21216)

	Ott18.fairseq.en-de.18.pretrained
de	BLEU = 30.39 61.4/36.0/23.8/16.2 (BP = 1.000 ratio = 1.018 hyp_len = 179739 ref_len = 176623)
en	BLEU = 35.05 66.2/41.7/29.0/20.8 (BP = 0.976 ratio = 0.976 hyp_len = 289499 ref_len = 296624)
fr	BLEU = 22.97 58.1/30.5/18.2/11.4 (BP = 0.934 ratio = 0.936 hyp_len = 67196 ref_len = 71785)
ru	BLEU = 20.34 53.0/25.7/15.0/9.1 (BP = 0.979 ratio = 0.980 hyp_len = 9138 ref_len = 9329)
es	BLEU = 20.97 55.2/27.3/15.8/9.7 (BP = 0.957 ratio = 0.958 hyp_len = 78384 ref_len = 81826)
cz	BLEU = 20.64 54.9/27.0/15.4/9.2 (BP = 0.965 ratio = 0.965 hyp_len = 41321 ref_len = 42812)
cs	BLEU = 17.86 52.3/24.4/13.4/7.8 (BP = 0.933 ratio = 0.935 hyp_len = 24419 ref_len = 26113)
it	BLEU = 19.08 51.4/24.2/13.6/7.8 (BP = 1.000 ratio = 1.010 hyp_len = 10327 ref_len = 10229)
hu	BLEU = 13.45 44.5/17.6/8.8/4.7 (BP = 1.000 ratio = 1.008 hyp_len = 21378 ref_len = 21216)

	Edunov18.fairseq.en-de.18.pretrained
de	BLEU = 38.54 67.4/44.2/31.7/23.4 (BP = 1.000 ratio = 1.004 hyp_len = 177263 ref_len = 176623)
en	BLEU = 34.85 66.0/41.7/29.2/21.1 (BP = 0.966 ratio = 0.967 hyp_len = 286734 ref_len = 296624)
fr	BLEU = 23.65 58.6/31.5/19.1/12.2 (BP = 0.923 ratio = 0.926 hyp_len = 66482 ref_len = 71785)
ru	BLEU = 21.50 54.5/27.3/16.3/9.9 (BP = 0.971 ratio = 0.972 hyp_len = 9064 ref_len = 9329)
es	BLEU = 21.66 56.0/28.4/16.7/10.4 (BP = 0.946 ratio = 0.947 hyp_len = 77493 ref_len = 81826)
cz	BLEU = 21.64 56.0/28.4/16.4/10.0 (BP = 0.957 ratio = 0.958 hyp_len = 41018 ref_len = 42812)
cs	BLEU = 19.31 53.8/26.1/14.9/8.9 (BP = 0.929 ratio = 0.932 hyp_len = 24333 ref_len = 26113)
it	BLEU = 19.68 52.2/25.1/14.0/8.2 (BP = 1.000 ratio = 1.005 hyp_len = 10280 ref_len = 10229)
hu	BLEU = 14.75 45.9/19.2/9.9/5.4 (BP = 1.000 ratio = 1.014 hyp_len = 21522 ref_len = 21216)

	Ng19.fairseq.en-de.19.pretrained
de	BLEU = 40.47 68.5/46.0/33.6/25.3 (BP = 1.000 ratio = 1.012 hyp_len = 178754 ref_len = 176623)
en	BLEU = 39.55 69.1/45.9/33.3/24.8 (BP = 0.983 ratio = 0.983 hyp_len = 291721 ref_len = 296624)
fr	BLEU = 25.82 59.9/33.4/20.8/13.7 (BP = 0.940 ratio = 0.941 hyp_len = 67576 ref_len = 71785)
ru	BLEU = 24.43 56.4/30.1/18.6/12.0 (BP = 0.984 ratio = 0.985 hyp_len = 9185 ref_len = 9329)
es	BLEU = 23.58 57.0/29.9/18.2/11.6 (BP = 0.962 ratio = 0.963 hyp_len = 78778 ref_len = 81826)
cz	BLEU = 22.86 56.5/29.3/17.3/10.7 (BP = 0.971 ratio = 0.972 hyp_len = 41600 ref_len = 42812)
cs	BLEU = 21.15 55.0/27.7/16.3/10.1 (BP = 0.944 ratio = 0.946 hyp_len = 24700 ref_len = 26113)
it	BLEU = 20.27 52.4/25.6/14.5/8.7 (BP = 1.000 ratio = 1.019 hyp_len = 10422 ref_len = 10229)
hu	BLEU = 14.42 45.3/18.8/9.7/5.2 (BP = 1.000 ratio = 1.039 hyp_len = 22053 ref_len = 21216)

	Wu19-dynamicglu.fairseq.en-de.16.pretrained
de	BLEU = 29.57 60.3/35.2/23.0/15.6 (BP = 1.000 ratio = 1.019 hyp_len = 179942 ref_len = 176623)
en	BLEU = 34.17 64.9/40.8/28.3/20.4 (BP = 0.972 ratio = 0.973 hyp_len = 288558 ref_len = 296624)
fr	BLEU = 22.88 57.6/30.3/18.0/11.2 (BP = 0.939 ratio = 0.941 hyp_len = 67556 ref_len = 71785)
ru	BLEU = 19.43 52.6/25.0/14.3/8.7 (BP = 0.967 ratio = 0.968 hyp_len = 9028 ref_len = 9329)
es	BLEU = 20.36 54.7/26.8/15.4/9.3 (BP = 0.951 ratio = 0.952 hyp_len = 77879 ref_len = 81826)
cz	BLEU = 20.78 54.8/27.1/15.5/9.2 (BP = 0.967 ratio = 0.968 hyp_len = 41426 ref_len = 42812)
cs	BLEU = 17.51 52.0/24.2/13.4/7.8 (BP = 0.919 ratio = 0.922 hyp_len = 24078 ref_len = 26113)
it	BLEU = 18.67 50.9/23.8/13.2/7.6 (BP = 0.999 ratio = 0.999 hyp_len = 10220 ref_len = 10229)
hu	BLEU = 13.40 44.9/18.0/9.0/4.9 (BP = 0.976 ratio = 0.976 hyp_len = 20705 ref_len = 21216)