Comparing Multilingual NMT Models and Pivoting

Celia Soler Uguet **Fred Bane** **Anna Zaretskaya** **Tània Blanch Miró**
TransPerfect
{csuguet,fbane,azaretskaya,tblanch}@translations.com

Abstract

Following recent advancements in multilingual machine translation at scale, our team carried out tests to compare the performance of pre-trained multilingual models (M2M-100 from Facebook and multilingual models from Helsinki-NLP) with a two-step translation process using English as a pivot language. Direct assessment by linguists rated translations produced by pivoting as consistently better than those obtained from multilingual models of similar size, while automated evaluation with COMET suggested relative performance was strongly impacted by domain and language family.

1 Background and Motivation

As a translation company, our work involves hundreds of distinct translation directions across dozens of languages. However, demand is not evenly distributed across all language pairs. The vast majority of our translation requests involve English as either the source or target language, with most other requests concentrated in a few major languages, such as German, French, Italian, Japanese, and Chinese.

Our fleet of machine translation (MT) engines is developed considering both the demand and the resources available for training. Currently, we use mostly bilingual models with some many-to-one models (such as for Scandinavian languages), but no many-to-many models. For language pairs where only a few hundred words are translated each year, the demand does not justify the costs incurred in training, deploying, and maintaining an engine for that language pair. Moreover, these language pairs often have scant high-quality resources available for training. Thus, in situations where demand for machine translation exists, but in insufficient amount to offset training and deployment costs, we have historically chosen to use a two-step translation process: pivoting through a related, high-resource language.

In recent years, multilingual models have shown growing potential to wholly or partially replace a fleet of bilingual models. The benefits are clear: no error propagation resulting from using the output of one model as the input to another as in the pivot scenario; reduced overhead and complexity by using one model for multiple language directions instead of separate models for each direction; improved translation quality in low-resource languages due to knowledge transfer from related languages; the potential for zero-shot translation for language directions for which no direct data exist, and so on. However, these models also have their drawbacks, including the expense and difficulty of retraining the models, the inability to add additional languages without retraining the model entirely, and the near impossibility of fine-tuning the model for particular clients.

Below we report the results of an experiment comparing bilingual base transformers (Vaswani et al., 2017) with pre-trained M2M-100 from Facebook obtained from Hugging Face (Fan et al., 2020) and multilingual models made public by Helsinki-NLP (Tiedemann and Thottingal, 2020),[1] using data drawn from our previous translation work and out-of-domain corpora.

[1] https://github.com/Helsinki-NLP/Opus-MT

2 Related Research

Interesting and very promising work has been carried out recently on multilingual MT approaches, where instead of training one NMT model for each language pair separately, a single model is trained that can translate from a single source into multiple target languages, or even many-to-many models that can translate in any direction between the languages they are trained on. Apart from improving MT performance for low-resource languages that can benefit from such models, these works also show competitive performance for resource-rich languages, suggesting the possibility of fully replacing the bilingual approach in the near future.

Most recently, the Facebook AI research group proposed a single multilingual translation model able to translate within any pair of the 100 languages included (Fan et al., 2020). The authors observed a significant improvement in performance in non-English language pairs, and a competitive performance in language pairs that include English compared to the WMT baseline from previous years (Barrault et al., 2019; Bojar et al., 2017; Bojar et al., 2018)

Multilingual MT models have been a subject of research for a few years now. In most cases, the goal has been to leverage parallel data available for resource-rich languages to improve MT performance for languages with scarce resources. As early as in 2015, Dong et al. (2015) explored an approach for simultaneously translating the same source sentence into multiple target sentences. They obtained a better performance on all language pairs (English into French, Spanish, Dutch and Portuguese) when using the multilingual model as opposed to single-target RNN models. However, statistical significance of the deltas are not indicated in the paper.

A few other works report significant improvement for low-resource languages thanks to multilingual models. Fira et al. (2016) propose a multi-way multilingual model trained on WMT'15 data. Ha et al. (2016) explore a multilingual NMT approach and report on promising results for low-resource languages, as well as in scenarios where there are not enough parallel data available in order to train a bilingual NMT model while achieving good performance.

A simpler multilingual NMT approach was proposed by Johnson et al. (2016). It does not require any change to the model architecture, but instead introduces a token at the beginning of the input sentence to indicate the target language. The authors report improvement for low-resource languages but, unlike the majority of other similar works, they observed a degradation on high-resource languages compared to bilingual models.

Finally, Tan et al. (2019) propose one more interesting approach, namely to use NMT with knowledge distillation, where bilingual models act as teachers. The authors report similar or improved results compared to the bilingual models used in the experiment.

It is notable that most of these works report very encouraging results: multilingual models always seem to outperform bilingual ones for low-resource languages, and perform en par or better for resource-rich languages. This contributes to the intuition that they will perform mostly better than two-level systems that pivot through English.

3 Materials and Methods

For this research, we set out to compare the performance of our company's pivoting system with open-source pre-trained multilingual models. For the pivoting system, we used general-purpose models trained to handle the different content types we have historically received in our translation work. These models were trained with between ten and thirty million sentence pairs, for fifty epochs or until the early stopping criterion was met (no improvement in validation set perplexity for 6 successive validation checkpoints). We used the transformer-base architecture with guided alignment using alignment from fast align (Dyer et al., 2013), and to limit potential confounding factors we use English as the pivot language for all language pairs. We chose to compare our system with two M2M-100 systems (the 480 million and the 1.2 billion parameter models) (Fan et al., 2020) and the multilingual models from Helsinki-NLP (Tiedemann and Thottingal, 2020). While there are other pre-trained multilingual SOTA models such as mT5 that could be fine-tuned for the downstream task of multilingual translation (Xue et al., 2021), we believe that the M2M-100 and Helsinki-NLP models were easily accessible and ready to be used with no further fine-tuning. Moreover, since all these systems were released around the same time, there is no published or reliable research to suggest that one model outperforms the rest.

We selected seven language pairs for which we received requests in the past year but for which we had no direct bilingual model. These were the following:

- Italian-French (referred to as IT–FR);
- French-Japanese (referred to as FR–JA);
- French-Chinese (referred to as FR–ZH);
- Spanish-Italian (referred to as ES–IT);
- French-Portuguese (referred to as FR–PT);
- Italian-German (referred to IT–DE);
- French-Arabic (referred to as FR–AR).

We also carried out a quantitative comparison for Danish–Spanish and Swedish–French, but since we could not find linguists available for the human evaluation, we do not include the results for these two pairs of languages.

In our experiments we used data from two different sources to avoid biases and compare performance across multiple domains. The first set of data was drawn from our company's previous human translation work (with care being taken to ensure that none of the data had been seen by the models during training). Although the data involved a wide variety of content types, we consider these test data to be "in domain" for our engines as they were sampled from essentially the same distribution as our training data. The second set of data was extracted from Leipzig's Corpora Collection (Goldhahn et al., 2012). This collection includes monolingual corpora for 291 Languages. Being a monolingual database, we can be quite confident that none of those texts were used for the training of any of the engines we were comparing. We extracted text from the news domain and from the most recent year available for each source language. These test data were considered to be "out-of-domain" for our engines.

Since no reference translations were available for any of the input sentences, we performed automated, reference-free evaluation using COMET, which was Unbabel's submission for the WMT 2020 Quality Estimation Shared Task (Rei et al., 2020). The reason behind this decision was that this model ended on the top 5 of best models in all tasks and language pairs but one. Moreover, it can be used for document-level assessment, it is easily accessible, it can be run on GPU, and it offers a command to compare multiple systems with statistical testing. Additionally, we also engaged human linguists to carry out blinded direct assessment (DA) for each language pair. Ordinarily we would commission multiple linguists for each language pair to mitigate the effects of bias and human error. However, for these less common translation directions, only one linguist was available per language pair. Nevertheless, we consider these scores reliable as the linguists were selected from our pool of certified translators for the language pair. This means that the annotators were not simply bilingual speakers, but held translation certification and actively performed translation tasks in this language pair.

Each linguist scored 200 segments chosen at random (100 from the in-domain data and 100 from the out-of-domain data) using a scale from 0 to 100. Linguists were instructed to score the segments based on the general quality of the MT output – how well it represented the main message of the input sentence – rather than small errors which would be more heavily penalized when evaluating human translations. The scoring criteria provided to the linguists were as below:

- 0: Completely unintelligible and useless translation;
- 25: Most of the target needs editing, but part of the MT can be preserved;
- 50: Half of the output is usable and half needs to be edited;
- 75: Edits needed, but MT output is usable;
- 100: Perfect translation, fully accurate.

Statistical significance for automated metrics was calculated using the bootstrap t-test from COMET (Koehn, 2004), and statistical significance for human DA was determined using unpaired t-test with p <0.01 considered statistically significant.

4 Results

The results of the human and automated evaluations are presented in Tables 1 through 3 below. In every case, human evaluation favored the translation from the pivot system, often by a large margin. This was true for both test sets as well as the overall scores. The difference was more pronounced for language pairs from different families than for language pairs where both the source and target were European languages (average difference of 10.99 in the overall scores for FR–JA, FR–ZH, and

FR–AR vs. 3.59 for IT–FR, ES–IT, FR–PT, and IT–DE).

COMET scores were less conclusive, suggesting that relative performance was more dependent on the domain of the content and the language families to which the source and target belonged. On the in-domain test set, scores for the pivot system were better than the small M2M-100 model in all but one language pair (FR–PT), and even outperformed the larger M2M-100 model in the three inter-language-family language pairs (FR–JA, FR–ZH, and FR–AR). For the European language pairs, the larger M2M-100 system obtained scores significantly higher than those for the pivot system.

For the out-of-domain test set, on the other hand, the M2M-100 models obtained higher scores in all language pairs, though we may again observe that scores for language pairs from different language families are roughly 50 percent lower than those for European language pairs.

4.1 Divergence Between COMET and DA Scores

In a number of instances, we noted pronounced divergence between the scores assigned by COMET and those from human linguists. To better understand this phenomenon, we manually analyzed some of these sentences and provide some examples in Table 4.

We find that in general those segments being given a low score by COMET but a higher score by human reviewers tend to contain a large number of punctuation marks, numbers, or proper nouns (especially those written in Latin characters when the language uses a different script). We speculate that low scores due to proper nouns may suggest a difference between COMET's linguistic knowledge and world knowledge, while the low scores for sentences in the former two categories may be related to the composition of the training data used to train the COMET system.

We present as well a comparison of the agreement between human reviewers and COMET. The plots for each language pair can be found from Figures 1 and 2 in Appendix A. X values represent the normalized difference in COMET scores between the M2M-100 translation and the translation of the pivot system; Y values represent the normalized difference in human scores respectively. Positive values represent a better score from the M2M-100 system, and negative values represent a better score from the pivoting system. Data points in quadrants I and III represent agreement between the human evaluation and COMET, while those in quadrants II and IV represent disagreement.

5 Discussion

In this study we compared translations from different models using human DA and automated evaluation with the COMET quality estimation model. We tested model performance using a combination of data sampled from the same distribution as our training data (in-domain) and news data (which were out-of-domain for our models used in the pivot system). Single-blind human DA showed a clear preference for the translations obtained through pivoting, while automated evaluation with COMET was less conclusive: the domain of the content and whether or not the source and target languages belonged to the same language family appeared to have a significant effect on the scores.

Beyond translation quality, as a translation company we must also take other aspects into consideration. While these fall outside the scope of this work, there are many other relevant factors, such as:

- Simplicity in production: It might be more desirable to have one model instead of many;
- Resource requirements: While one model can take the place of many, multiple instances of the model would be needed, and each instance requires greater resources, so the ultimate effect on hosting and inference costs is uncertain;
- Updating problems: With a multilingual model it is more complex and costly to update or fix problems that are discovered during inference. It is much easier to retrain bilingual models in response to issues;
- Adding more languages: It is not possible to add more languages to an already-trained multilingual model, whereas a pivoted approach can be deployed on-demand for any two languages that are supported with bilingual models;
- Client customization: It is unclear how, if at all, a multilingual model may be adapted for particular clients, especially clients with

	Overall			In-Domain			Out-Of-Domain		
Lg. Pair	Pivot	M2M	Helsinki	Pivot	M2M	Helsinki	Pivot	M2M	Helsinki
IT–FR	**73.64**	68.35	64.66	**70.04**	67.25	64.54	**76.89**	69.42	64.77
FR–JA	**69.86***	58.84	N/A	**71.34***	56.15	N/A	**68.45***	61.4	N/A
FR–ZH	**73.18***	65.56	N/A	**78.23***	66.56	N/A	**68.07**	64.56	N/A
ES–IT	**83.3**	78.98	76.02	**88.3**	81.53	79.2	**78.3**	76.43	72.9
FR–PT	**90.63**	88.21	84.59	**91.79**	87.78	83.23	**89.47**	88.65	85.94
IT–DE	**86.2**	83.85	N/A†	**78.95**	76.58	N/A†	**93.68**	91.28	N/A†
FR–AR	**67.8**	53.46	N/A	**76.73**	51.72	N/A	**58.86**	55.2	N/A

Table 1: Human direct assessment scores for each system. The M2M-100 system used here is the smaller of the two (480M), so as to be directly comparable with the base transformers used in the pivot system. * Indicates scores with a statistically significant difference ($p < 0.01$). †Indicates that no multilingual model was available, only a direct bilingual model.

Language Pair	Pivot	M2M (480M)	M2M (1.2B)	Helsinki-NLP
IT–FR	0.3773	0.3608	**0.4035***	0.3216
FR–JA	**0.2305**	0.1937	0.2222	N/A
FR–ZH	**0.1944**	0.1563	0.1728	N/A
ES–IT	0.4704	0.4464	**0.4877***	0.3903
FR–PT	0.3711	0.3782	**0.4026***	0.3372
IT–DE	0.3271	0.2901	**0.3498***	N/A†
FR–AR	**0.2003**	0.1875	0.1574	N/A

Table 2: COMET scores for each system on in-domain data. * Indicates scores with a statistically significant improvement compared to the Pivot column ($p < 0.01$). †Indicates that no multilingual model was available, only a direct bilingual model.

Language Pair	Pivot	M2M (480M)	M2M (1.2B)	Helsinki-NLP
IT–FR	0.3158	0.3223	**0.3934***	0.2698
FR–JA	0.1816	0.1889	**0.227***	N/A
FR–ZH	0.1376	0.1401	**0.1783***	N/A
ES–IT	0.3771	0.3987*	**0.4487***	0.3418
FR–PT	0.3394	0.4042*	**0.4543***	0.3395
IT–DE	0.2302	0.229	**0.3158***	N/A†
FR–AR	0.1943	**0.2141***	0.1751	N/A

Table 3: COMET scores for each system on out-of-domain data. * Indicates scores with a statistically significant improvement compared to the Pivot column ($p < 0.01$). †Indicates that no multilingual model was available, only a direct bilingual model.

Language Pair	Source	Target	COMET[2]	Linguist
IT–FR	La siringa contiene <<mL COUNT>> ml di soluzione iniettabile, da <> mg <>, <> mg <> o placebo.	La seringue contient <<mL COUNT>> ml de solution injectable, de <> mg <>, <> mg <> ou placebo.	27.69	100
FR–JA	C’est une rentrée pleine d’incertitudes à l’hôpital , confirme Mélanie Meier, de la CFDT.	「これは不確実性に満ちた病院への帰還だ」とCFDTのメラニー・メイエ氏は述べている。	0	80
FR–ZH	Je travaille pendant les vacances à Dour et à Pukkelpop et j’ai normalement beaucoup d’argent de poche l’été.	我在Dour和Pukkelpop度假期工作,我通常在夏天有很多。	0	90
ES–IT	jersey de rayas anchas con cuello a la caja.	Maglia a righe larghe con scollo.	0	90
FR–PT	Tribunal de Paris – Corruption : Après Lamine Diack, Papa Massata condamné. . .	Tribunal de Paris – Corrupção: depois de Lamine Diack, Papa Massata condenada...	29.34	99
IT–DE	2.2 Come meglio descritto nel dettaglio al successivo art.	2.2 Wie besser in der Kunst ausführlich beschrieben.	0	98
FR–AR	CHRU DE LILLE - Hôpital Albert Calmette	CHRU DE LILLE - مستشفى ألبرت كالميت	20.30	90

Table 4: Some examples of segments with a low COMET score in comparison to the score given by the linguist.

small translation memories or those who translate in only one language pair;

- Trade-off between low- and high-resource languages: Performance in low-resource languages can be improved through knowledge transfer from higher-resource languages, but decreased performance in these higher-resource languages may outweigh these gains due to the greater volume of demand.

Contrary to our intuitions prior to undertaking this study, our results suggest that pivoting is a reasonable choice for language pairs where no direct model exists, at least in terms of translation quality. The strength of the conclusions are limited by the relatively small sample size, and we anticipate these results will need to be revisited as multilingual models become more capable. Moreover, fine-tuning other pre-trained multilingual models such as mT5 and comparing those with the pivoting approach could lead to different conclusions. Further research is needed to more comprehensively weigh the advantages and disadvantages of replacing multiple bilingual models with a single multilingual model.

References

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. 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)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Bojar, Ondrej, Chatterjee Rajen, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, and Christof et al. Monz. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Second Conference on Machine Translation*, pages 169–214. The Association for Computational Linguistics.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July. Association for Computational Linguistics.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. 10.

Firat, Orhan, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*, abs/1601.01073.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Ha, Thanh-Le, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

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

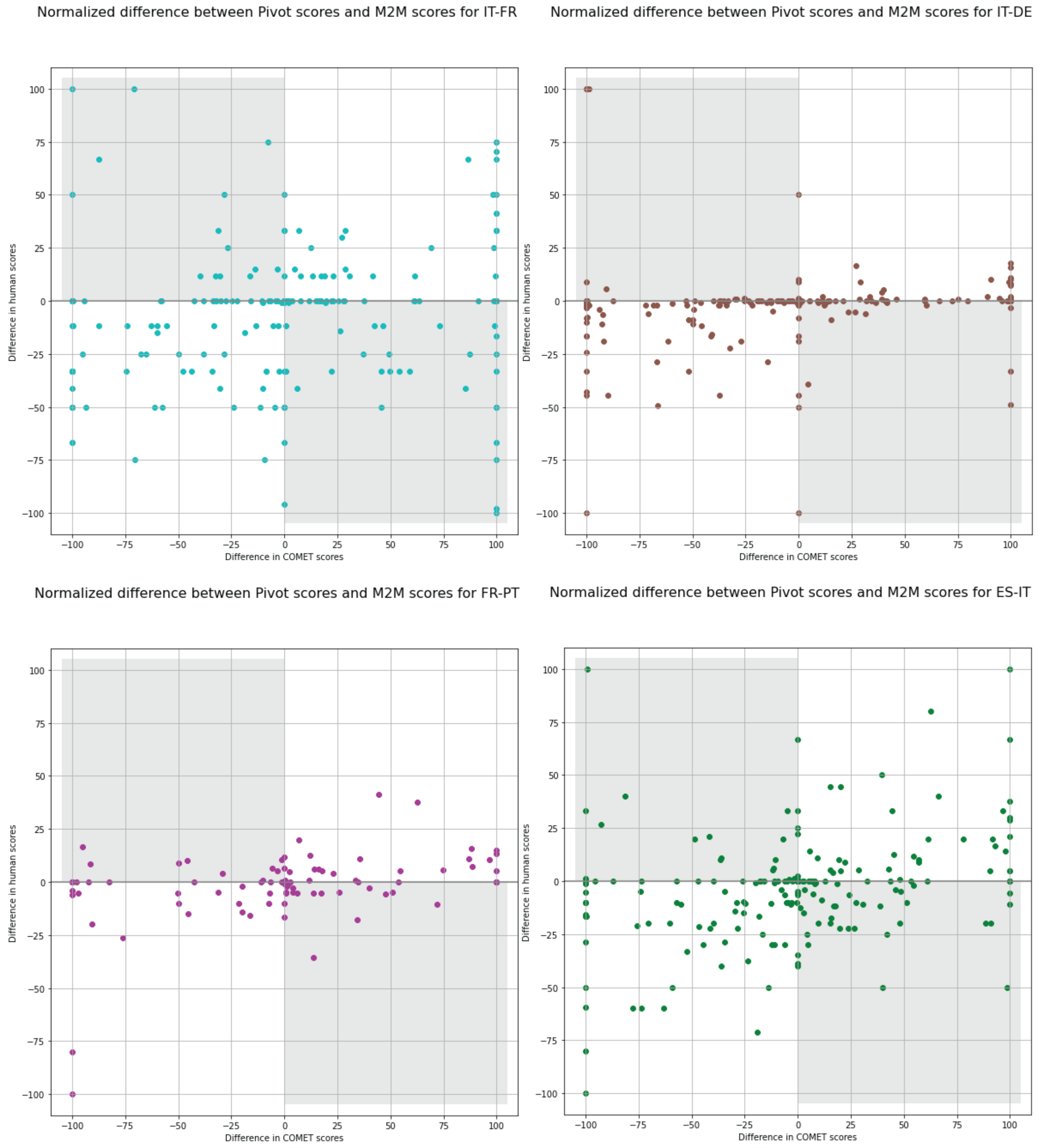
Tan, Xu, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *CoRR*, abs/1902.10461.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Appendix A. Comparison of COMET and Human DA Scores



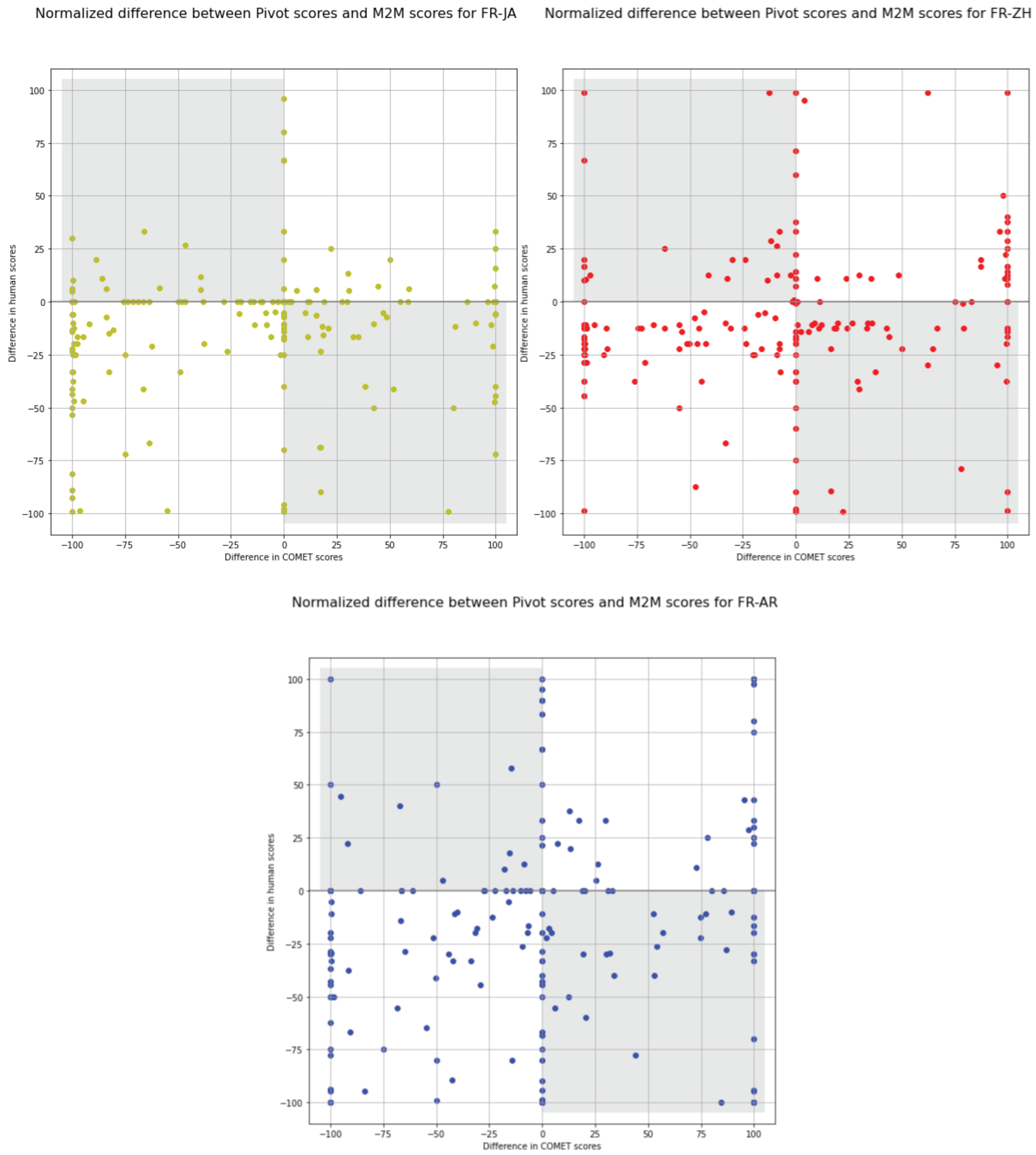
Normalized difference between Pivot scores and M2M scores for IT-FR

Normalized difference between Pivot scores and M2M scores for IT-DE

Normalized difference between Pivot scores and M2M scores for FR-PT

Normalized difference between Pivot scores and M2M scores for ES-IT

Figure 1: Comparison of difference between COMET and human annotations: language pairs in the same language family



Normalized difference between Pivot scores and M2M scores for FR-JA

Normalized difference between Pivot scores and M2M scores for FR-ZH

Normalized difference between Pivot scores and M2M scores for FR-AR

Figure 2: Comparison of difference between COMET and human annotations: language pairs in different language families