Using Machine Translation to Augment Multilingual Classification

Adam King
GumGum
aking@gumgum.com

Abstract

An all-too-present bottleneck for text classification model development is the need to annotate training data and this need is multiplied for multilingual classifiers. Fortunately, contemporary machine translation models are both easily accessible and have dependable translation quality, making it possible to translate labeled training data from one language into another. Here, we explore the effects of using machine translation to fine-tune a multilingual model for a classification task across multiple languages. We also investigate the benefits of using a novel technique, originally proposed in the field of image captioning, to account for potential negative effects of tuning models on translated data. We show that translated data are of sufficient quality to tune multilingual classifiers and that this novel loss technique is able to offer some improvement over models tuned without it.

1 Introduction

One of the most common uses of machine learning for natural language processing (NLP) is the classification of text into one of multiple mutually-inclusive or mutually-exclusive labels. Recently, generative LLMs, such as PaLM (Chung et al., 2022) and ChatGPT (Ouyang et al., 2022) have shown exciting and impressive capabilities to do zero- or few-shot prompting, classify text given only a few examples for the task across a variety of languages. Nevertheless, it is still the case that the highest performing and most efficient means to classify text is the use of a bespoke classifier trained with hundreds or thousands labeled examples (Pires et al., 2019), particularly when the task requires a level of human-like subjectivity or general reasoning ability (Kocoń et al., 2023, see discussion). To this end, finding or creating a corpus of labeled examples is a necessary step in the creation of any classifier.

For high-resource languages like English, which have many existing labeled corpora available and large populations of annotators on crowd-sourced workers such as Amazon Mechanical Turk, the challenge of creating or finding training and evaluation data can be costly, but not prohibitively so. Yet, for lower-resourced languages which lack existing annotated corpora and have smaller or even non-existent populations on these large annotation platforms, acquiring the required training data can prove to be much more difficult. Moreover, if the model is intended to be able to perform the same classification across multiple languages, the time and effort required to annotate training data becomes multiplicative. Fortunately, classification is not alone in the applications of machine learning in NLP. Machine translation (MT) has seen major improvements in recent years (Stahlberg, 2020), accelerated by the adoption of the transformer architecture (Vaswani et al., 2017).

To date, several options for high quality machine translation currently exist, between API services and open-source models. MT API services, such as Google translate, have become nearly ubiquitous, provide high quality translations, while still being relatively inexpensive. In fact, in one experiment, translating data using Google translate into English and using existing English-trained classifier models outperformed certain models trained on the original language directly (Araujo et al., 2016). In addition to MT API services, several open-source translation models are easily available, such as the multilingual M2M100 model (Fan et al., 2020), NLLB200 model (Team et al., 2022) or the over 1400 models trained by the University of Helsinki (Tiedemann and Thottingal, 2020), with many of these models have performance that approaches or exceeds that of MT APIs (Stahlberg, 2020).

With this in mind, it may be the case that translating an existing, labeled dataset with one of the aforementioned MT options is a feasible alternative to creating a novel dataset directly in that language. This has several benefits. Firstly, it avoids the problem of existing corpora or annotation options not existing for the language in question. Secondly, it minimizes the data needed for multilingual models and allows annotations for one language to serve another. Here, we ask if it is possible to use MT to train a multilingual model, given only original, annotated data for a single language.

Of course, the potential benefits of using MT to train a multilingual model are still affected by the old machine learning adage: garbage in, garbage out. Even the best translations, either human or machine, will lose some of the information of the original language, which will inevitably lead to dropped performance for a model trained on the translated examples. Fortunately, the problem of training models using semantically similar but imperfect pairs of data is not unique to the task at hand and there is a growing body of research which may provide some benefit. In particular, image captioning is a task to generate the ideal natural language text caption for an image and these captioning models must learn to represent semantically related data from very different modalities similarly, i.e., text and images (Li et al., 2021). In this way, image captioning is somewhat analogous to the task of training on translated data, where we want to have semantically identical text from different languages predicted to have the same labels. As a result, we ask in addition whether some of the model training techniques used in image captioning models can lead to improved performance for multilingual models trained using MT data.

2 Related Work

This work is by no means the first to suggest the usage of machine translation to create or augment datasets for lower resourced languages. Wei and Pal (2010) and Pan et al. (2011) augmented Chinese language corpora with annotated data translated from English to improve the performance of a Chinese-language sentiment analysis model. On the other hand, Barriere and Balahur (2020) and Ghafoor et al. (2021) used existing API translation services to translate annotated data from English into lower-resourced languages and trained classifiers solely on these translated data, finding that classifiers trained on translated data were fairly accurate but did see drops in performance, likely due to the effects of imperfect translations of the training data.

It should be noted that training a model from scratch is not the only means to create an accurate classifier, particularly for lower-resourced languages. Large multilingual transformer models such as M-BERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019) or GPT-3 (Brown et al., 2020) have been shown to have the ability to generalize from one language to the other, i.e., train in one language and improve test performance in another language, (Pires et al., 2019), but benefits of this vary on the languages in question, with languages that share closer genealogical origin or structural similarities benefiting more from inter-language transfer. Regardless, training a model with examples of a particular language dependably yields the best classifier for new data in that language.

Nevertheless, to date there has been no investigation of how fine-tuning large multilingual transformer models on translated data affects final performance compared to simple interlanguage transfer. Moreover, previous work to train models using translated data employed a naive approach, treating translated data as if it were no different than original, untranslated data which annotated itself. In this work, we investigate both how multilingual transformer models trained on translated data perform compared to interlanguage transfer and explore a means to mitigate imperfect translation quality when creating these training datasets.

3 Image captioning and Image-Text Contrastive Loss

Image-text Contrastive (ITC) loss is a technique used when training multimodal models to caption images with natural language descriptions (Li et al., 2021). For example, BLIP (Li et al., 2022) is a image-captioning model that was trained with a mix of human- and artificially-annotated images where ITC loss was integral to the models ability to learn from noisy, artificially-annotated data. ITC loss, then, has been shown to mitigate negative effects of both noise and different modalities for multimodal models.

At an intuitional level, these captioning models decompose text and images into a shared embedding space and ITC loss seeks to penalize cases where related image-text pairs are dissimilar in this shared embedding space. In other words, ITC looks seeks to bring semantically related items from disparate modalities closer in a shared embedded space and has empirically improved image-captioning models, with little impact on training time or resources.

Training multilingual classification models with translated data bears a similarity to captioning, though rather than have semantically related examples from different modalities, there are semantically parallel data in different languages. That being the case, we will be a slightly modified form of ITC loss, namely original-translated contrastive (OTC) loss, to enforce similarity within a batch between data from the original language and its translated counterpart. Like ITC loss, OTC loss penalizes a transformer model for dissimilar embedding representations for translated pairs. One way to think of it is that this loss encourages the model to embed sentences with the same meaning identically, regardless of language.

In detail, we implement OTC loss as follows. We begin by deriving a probability of each original/translated pairing in a training minibatch, p^{o2t} and p^{t2o}, that is, which original examples pairs with which translated example and vice versa.

$$p_m^{o2t} = \frac{exp(s(O, T_m)/\tau)}{\Sigma_m^M exp(s(O, T_m)/\tau)} \tag{1}$$

$$p_m^{t2o} = \frac{exp(s(T, O_m)/\tau)}{\Sigma_m^M exp(s(T, O_m)/\tau)} \tag{2}$$

Here, $s(T, O)$ is a similarity function between the original, untranslated data and the translated examples in a minibatch. We compute $s(T, O)$ by first extracting and normalizing the embedding for the initial [CLS] token after the final attention head of the encoder stack in M-BERT, computing a pairwise dot product for all possible pairs of original and translated data and dividing by τ, which is a learnable parameter. We then apply the softmax function as a way to represent the likelihood of each original/translated match. Ideally, each correct original/translated pair will have the most similar embeddings, resulting in a value close to 1 after softmax. As a final step, we compute the cross-entropy between the result of the previous step and a target vector which encodes the correct original/translated pairs, weighting this by a hyperparameter, α_{otc}. Following BLIP (Li et al., 2022), we set $\alpha_{otc} = .4$ for all runs.

$$\ell_{otc} = \alpha_{otc} * \frac{1}{2}\mathbb{E}_{(O,T)}[H(\mathbf{y}^{o2t}(O), \mathbf{p}^{o2t}(O) + H(\mathbf{y}^{t2o}(T), \mathbf{p}^{t2o}(T)] \tag{3}$$

4 Experiments

4.1 Data

For these experiments, we use a multilingual dataset of Amazon product reviews across 6 languages: English, Spanish, French, German, Chinese and Japanese (Keung et al., 2020). This dataset is comprised of over 1 million total examples, split into a train and test partition. The reviews are equally distributed across the six languages, as well as the total stars given to the reviewed product (1-5) for both the train and test partition, i.e., each number of stars comprises 20% of the examples for that language. This dataset is particularly useful due to its size, number of available languages and presence of an established training and test data split.

We began by translating each review from the training partition of the original dataset into each of the other respective languages and assigned the same star value to the review (see example 1), i.e., if a review was originally in English and had star star, when translating it into French it would also be labeled with one star. We did this translation once before carrying out the rest of the experiment to ensure each classifier would be trained on the same set of translations. To translate, we used a single multilingual translation model, M2M100 (Fan et al., 2020). We chose to use a single multilingual translation model in order to mitigate any potential differences from translation quality coming from different machine translation architectures.

4.2 Experiment design

To investigate any potential improvement in classifier accuracy with the use OTC loss, we fine-tuned

id	translated	language	text	stars
1	0	en	My daughter really likes the backpack and ...	5
1	1	es	Mi hija realmente le gusta el bolsillo y ...	5
...	...	...	...	...
2	0	en	This product is BS, I washed my face with hot water ...	1
2	1	fr	Ce produit est BS, je me suis lavé le visage à l'eau chaude ...	1
...	...	...	...	...

Figure 1: Example original and translated data. Each unique review (id) in the original dataset was translated to the other languages and assigned the same star value. Texts truncated here for formatting.

pretrained transformer models on datasets that included original, untranslated data for a single language[1] and only translated data for all others in the six language set. As an example, in one training run, the model would be tuned on the original English training data and only translated data for all other languages, which were translated from the set of the original English data. We did this for all six languages in the original set to ensure any results were not restricted to one language in the dataset. Though the exact training examples varied for each model, we tested each on the original testing split of the dataset, which was solely comprised of original data, i.e., non-translated, for the six languages.

In each case, we tuned a multilingual DISTILBERT model (Sanh et al., 2019), a distilled version of the original multilingual M-BERT (Devlin et al., 2018), to predict the number of stars on a review as a categorical classification problem, using categorical cross-entropy loss and varying between using OTC loss as an additional loss parameter between runs. We chose to use a distilled variant of BERT due to the distilled variants increased speed of training, while still maintaining 97% of overall language understanding of the original.

Because of the mechanics of OTC loss, each translated datum must have an original match in the minibatch and each original must have at least one translated variant. As such, we constructed minibatches during training such that half the samples were always original, untranslated data and the other half were a randomly selected translated example for each original datum. For each original example, we randomly selected a translated example from the other languages, meaning that the model saw an equal number of original and translated examples during tuning overall, though it saw far fewer individual examples of each translated language, i.e., roughly $\frac{1}{5}$. For simplicity, we restricted our tests to a 1:1 original:translated ratio and we used the same batch sampling method for runs without OTC loss, to make results more easily comparable.

For each tuning run, we used a batch size of 32 (16 original and 16 translated examples per batch)[2] and used the AdamW (Loshchilov and Hutter, 2017) optimizer with a linear warm-up of 500 updates with a learning rate of 2e-5. All training was done on G5.2XLARGE AWS instances which contain NVIDIA A10G GPUs. We tuned 3 separate tuning runs for each set of hyperparameters and report their mean values in the next section.

5 Results

In these experiments, we asked two simple questions: 1) how feasible is it to tune a multilingual transformer model on translated data and 2) does the inclusion of OTC loss improve model performance for languages where only translated training data was used.

In answer to the first, for each of the six languages in the original dataset, models finetuned with translated data showed higher F1-micro scores[3] on the held-out test set, compared to models trained with only original data for a single language (see Table 1). As was expected from Pires et al. (2019), even if a model was never exposed to data for a language, original or translated, the final model did have F1-micro greater than chance for that language (which would be 20% for a balanced, 5-label problem), indicating

[1] We restricted the experimental conditions to only including a single language's original data, rather than use the full set of $6! = 720$ possible permutations of language combinations for the sake of efficiency and resources.

[2] For baseline conditions where there was no translated data, mini-batching happened as normal with 32 examples original, untranslated data per batch.

[3] F1-micro is an example-weighted version of the F1-score, which is the harmonic mean or precision and recall. For more details on F1-score, see (Jurafsky and Martin, 2008).

Language	F1-micro		
	No data	Translated	Original
EN	0.407	0.481	0.554
FR	0.379	0.468	0.544
DE	0.359	0.465	0.581
ES	0.376	0.474	0.55
JA	0.307	0.396	0.543
ZH	0.352	0.372	0.458

Table 1: F1-micro for models trained with no samples for the specified language (No data), with only translated samples (Translated) and with the original training data for that language (Original). All languages saw a sizeable boost to performance over their respective baselines when using translated data (.02-.11) but all languages did perform markedly better when given actual data for each language.

Language	F1-micro	
	No OTC	OTC
EN	0.479	**0.483**
FR	0.464	**0.472**
DE	0.463	**0.467**
ES	0.472	**0.476**
JA	0.393	**0.399**
ZH	0.368	**0.376**

Table 2: Comparison on final performance per language for models that only included translated examples for the specified language. Though the gain was less than .1, each language consistently performed better when trained with OTC loss.

there was interlingual knowledge transfer happening within the model during training. Moreover, it appears that there was more transfer between related, similar languages, compared to more dissimilar languages; models trained with data for a European language showed higher performance on other European languages, compared to Japanese or Chinese. Nevertheless, for all languages, the use of translated data did show a noticeable improvement (.02-.11), though for each language, models trained with only translated data did underperform models trained with the full set of origina, untranslated training examples for that language (.07-.12).

That said, it is clear that the use of translated training data does improve model performance, even if the trained model only sees translated examples for that language. It should also be noted that due to the batching and sampling strategy used here, models trained with translated data saw far fewer examples of each language where they only saw translated data. That is, because each original review was paired with a single translated example out of five possible translated, these models were exposed to roughly one fifth of the data for translated languages and still saw a sizable boost in performance.

Moving on to the effect of OTC loss, Table 3 shows the mean F1-micro per language in the testing set, for models fine-tuned using original data for the specified language and translations for all other languages. For all languages, models trained using OTC loss saw an improvement over models trained without for all languages except Chinese, which showed a mixed set of negligible differences or lowered performance. However, these values include runs where the specific language was included as original, untranslated data. When averaging across all runs where a language in the testing set was only represented by translated data, OTC loss shows an improvement over models trained without it for all languages. Table 2 shows the mean F1-micro for all models trained where the specified language was not the original language.

To ensure that the results here were in fact statistically significant, we fit a linear mixed-effect model to predict final model F1-micro for a language, given the hyperparameters of a particular tuning run. Mixed-effect models are able to accurately evaluate the contribution of different fixed-effect independent variables, e.g., whether OTC was used when training a particular model, on dependent variables, e.g., the final accuracy of the trained model, all the while being robust to expected random variance between trials, e.g., because of random initialization and batching, some deep learning models score higher than others with identical hyperparameters (see Baayen et al. (2008), Jaeger (2008) for more).

This statistical model was fit to predict per-language test f1-micro, given a random effect of each model run and three fixed effects: i) the tested language, ii) the identity of the single original language and iii) whether OTC loss was added. OTC was found to have a significant, positive effect (COEF=0.036, STD.ERROR=0.017, for all model details see 2), indicating that even after taking into consideration differences between languages and random variance for each multilingual model, the inclusion of OTC loss did yield an improved final model F1-micro.

Orig. Training Language	OTC	EN	FR	DE	ES	JA	ZH
EN	No OTC	0.548	0.488	0.493	0.489	0.425	0.423
	OTC	**0.553**	**0.507**	**0.522**	**0.512**	**0.434**	0.422
FR	No OTC	0.504	0.539	0.504	0.493	0.424	**0.426**
	OTC	**0.512**	0.539	**0.517**	**0.511**	**0.428**	0.412
DE	No OTC	0.514	0.495	0.577	0.495	0.436	0.427
	OTC	**0.524**	**0.506**	**0.581**	**0.506**	**0.449**	0.425
ES	No OTC	0.506	0.497	0.500	0.544	0.433	**0.419**
	OTC	**0.523**	**0.510**	**0.518**	**0.548**	**0.441**	0.413
JA	No OTC	0.470	0.460	0.477	0.468	**0.526**	**0.436**
	OTC	**0.493**	**0.474**	**0.499**	**0.487**	0.522	0.424
ZH	No OTC	0.486	0.439	0.441	0.444	0.398	0.482
	OTC	**0.488**	**0.467**	**0.473**	**0.472**	**0.421**	**0.503**

Table 3: F1-micro results on untranslated test data. Each row shows the per-language performance for models trained with original data for the specified language and translated data for all other languages, using OTC loss and without. Each cell shows the mean of 3 runs per condition. Bolded values show a difference of .03 or greater.

6 Discussion and future directions

We investigated the feasibility of using translated text to fine-tune a multilingual transformer model, as well as any potential gains by utilizing a novel application of deep learning technique to improve performance. We found that models trained using only translated data for a language do show a noticeable improvement over baselines, though as expected, there was still a performance drop from using original, untranslated data for that language. We also found that slight further gains can be achieved by the use of OTC loss, suggesting that training the model in such a way where it is sensitive to potential data issues improves its ability to generalize.

Granted, this is a very open problem and results of using translated data to tune a multilingual classifier will vary highly depending on the quality of MT model used, architecture of the classifier being tuned and the type of classification being modeled. Nevertheless, the results here are exciting for multiple reasons. Firstly, as suggested by previous works (Shalunts et al., 2016, as an example), MT is useful tool for language-specific dataset creation when creating a dataset for that language directly may prove difficult. In this case, we showed that M-BERT models tuned on translated examples showed large gains over simple multilingual transfer during training. This is particularly interesting given that for each translated language, the model was only given a fraction of samples compared to the original language due to the 1:1 ratio of original and translated data. A future direction for this work may be to adjust this ratio or the number of languages in the dataset to investigate how this affects model training. Secondly, the use of OTC loss was shown to lead to a small, but robust boost to performance. This suggests that methods of mitigating the natural effects of translation have a potential to bridge the gap, so to speak, between models trained on translated data and on datasets in the target language directly. Particularly relevant, Chinese, which is linguistically dissimilar from the majority of languages in the set used here, showed a mixed ability to benefit from training with other languages, but a clearer improvement using OTC loss. This may suggest that OTC loss is able to mitigate structural differences between languages and a future direction for this may be to explore exactly how OTC loss affects individual examples and how other noise-reduction techniques may lead to further gains in model performance.

Putting this together, this is an indication that MT-augmented datasets stand as a good first step for developing multilingual classification models. Given that MT can quickly and efficiently expand an annotated dataset from one language into another and that translated dataset is of sufficient quality to improve over basic interlingual transfer, this technique has great potential to expanding classification tasks to new languages quickly. In addition, OTC loss may be able to slightly but significantly increase the quality of these models with no additional data. All in all, we are confident that the use of MT augmentation is an exciting and interesting topic for future exploration.

7 Acknowledgements

This work was carried out as a part of the R&D for GumGum's Verity product. Special thanks go to the members of the AI team (names, names, names) for their suggestions and audience during brainstorming, development and analysis of this project.

References

Araujo, Matheus, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st annual ACM symposium on applied computing*, pages 1140–1145.

Baayen, R Harald, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Barriere, Valentin and Alexandra Balahur. 2020. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. *arXiv preprint arXiv:2010.03486*.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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. *CoRR*, abs/2010.11125.

Ghafoor, Abdul, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, Mudasir Ahmad Wani, et al. 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490.

Jaeger, T Florian. 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446.

Jurafsky, Daniel and James H Martin. 2008. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Keung, Phillip, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

Li, Junnan, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651.

Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Loshchilov, Ilya and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pan, Junfeng, Gui-Rong Xue, Yong Yu, and Yang Wang. 2011. Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *Advances in Knowledge Discovery and Data Mining: 15th Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings, Part I 15*, pages 289–300. Springer.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Shalunts, Gayane, Gerhard Backfried, and Nicolas Commeignes. 2016. The impact of machine translation on sentiment analysis. *Data Analytics*, 63:51–56.

Stahlberg, Felix. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

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

Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, Bin and Christopher Pal. 2010. Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the ACL 2010 conference short papers*, pages 258–262.

```
              Mixed Linear Model Regression Results
==========================================================
Model:              MixedLM  Dependent Variable:  test_acc
No. Observations:   108      Method:              REML
No. Groups:         18       Scale:               0.0038
Min. group size:    6        Log-Likelihood:      111.7634
Max. group size:    6        Converged:           Yes
Mean group size:    6.0
----------------------------------------------------------
                       Coef.  Std.Err.   z    P>|z| [0.025 0.975]
----------------------------------------------------------
Intercept               0.465    0.024 19.128 0.000  0.418  0.513
otc[T.True]             0.036    0.017  2.105 0.035  0.002  0.069
original_lang[T.en]    -0.020    0.028 -0.714 0.475 -0.074  0.035
original_lang[T.es]    -0.012    0.028 -0.432 0.666 -0.066  0.042
original_lang[T.fr]    -0.015    0.028 -0.555 0.579 -0.070  0.039
original_lang[T.ja]    -0.030    0.028 -1.093 0.274 -0.085  0.024
original_lang[T.zh]    -0.050    0.028 -1.801 0.072 -0.104  0.004
test_lang[T.en]         0.016    0.020  0.762 0.446 -0.025  0.056
test_lang[T.es]         0.003    0.020  0.130 0.897 -0.037  0.043
test_lang[T.fr]        -0.000    0.020 -0.005 0.996 -0.040  0.040
test_lang[T.ja]        -0.061    0.020 -2.972 0.003 -0.101 -0.021
test_lang[T.zh]        -0.068    0.020 -3.336 0.001 -0.108 -0.028
Group Var               0.001    0.009
==========================================================
```

Figure 2: Full model details for MLE model trained to predict F1-micro per laguage. OTC has a positive contribution to an increase F1-micro score, even when controlling for variance between languages and model runs.