Zero-shot Cross-lingual Transfer is Under-specified Optimization

Anonymous ACL submission

Abstract

Pretrained multilingual encoders enable zero-shot cross-lingual transfer performance, but often produce unreliable models that exhibit high performance variance on the target language. We postulate that high variance results from *zero-shot cross-lingual transfer solving an under-specified optimization problem*. We show that the source language monolingual model and source + target bilingual model are linearly connected using a model interpolation, suggesting that the model struggles to identify good solutions for both source and target languages using the source language alone.

1 Introduction

Pretrained multilingual encoders like Multilingual BERT (mBERT; Devlin et al., 2019) and XLM-RoBERTa (XLM-R; Conneau et al., 2020) facilitate zero-shot cross-lingual transfer (Wu and Dredze, 2019; Hu et al., 2020) — training the model on one language then using it on another language without additional task-specific training data. While many have touted zero-shot successes with these models, the truth is that the outcome from any one experiment is highly variable. The choice of random seed makes the performance on a target language with cross-lingual transfer highly variable (Keung et al., 2020; Wu and Dredze, 2020) and makes it difficult to compare different models in the literature. Similarly, pretrained monolingual encoders also have unstable performance during fine-tuning (Devlin et al., 2019; Phang et al., 2018).

Why are these models so sensitive to the random seed? Many theories have bee offered: catastrophic forgetting of the pretrained task (Phang et al., 2018; Lee et al., 2020; Keung et al., 2020), small data size (Devlin et al., 2019), impact of random seed on task-specific layer initialization and data ordering (Dodge et al., 2020), the Adam optimizer without bias correction (Mosbach et al., 2021; Zhang et al., 2021), and a different generalization error with similar training loss (Mosbach et al., 2021). However, none of these factors fully explain the high variance of zero-shot cross-lingual transfer.

We offer a new explanation for high variance in target language performance: *the zero-shot cross-lingual transfer optimization problem is under-specified*. We hypothesize that these models have many good solutions when considering source language task data. However, these solutions perform differently on the target language, and only a small subset of them perform on par with models with target language supervision. Without target language supervision, the optimization is under-specified: you do not know what solution you will get.

Based on a linear interpolation of 1-dimensional plot and contour plot (Goodfellow et al., 2014; Li et al., 2018), we show that the monolingual source model and bilingual source and target model are linearly connected on the source language generation error surface. For the target language generation error surface, the performance increases smoothly as we move from a monolingual model to a bilingual model. This finding suggests that only a small subset of the solution space for the source language solves the target language; the optimization is unlikely to find such a solution without target language supervision, hence an under-specified optimization problem. By comparing both mBERT and XLM-R, we find that the generation error surface of XLM-R is flatter than mBERT, contributing to its better performance compared to mBERT.

2 Existing Hypotheses

Prior studies have observed encoder model instability, and have offered various hypotheses to explain this behavior. Catastrophic forgetting – when neural networks trained on one task forget that task after training on a second task (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017) —has been credited as the source of high variance in both monolingual fine-tuning (Phang et al., 2018; Lee et al., 2020) and zero-shot cross-lingual transfer (Keung et al., 2020). Mosbach et al. (2021) wonder why preserving cloze capability is important. In zero-shot cross-lingual transfer, deliberately preserving the multilingual cloze capability with regularization improves performance but does not eliminate the zero-shot transfer gap (Aghajanyan et al., 2021; Liu et al., 2021).

Small training data size often seems to higher variance in performance (Devlin et al., 2019), but Mosbach et al. (2021) found that when controlling the number of gradient updates, smaller data size has the similar variance as larger data size.

In the pretraining-then-fine-tune paradigm, random seeds mainly impact the initialization of task-specific layers and data ordering during fine-tuning. Dodge et al. (2020) show development set performance has high variance with respect to seeds. Additionally, Adam optimizer without bias correction—an Adam (Kingma and Ba, 2014) variant (inadvertently) introduced by the implementation of Devlin et al. (2019)—has been identified as the source of high variance during monolingual fine-tuning (Mosbach et al., 2021; Zhang et al., 2021). However, in zero-shot cross-lingual transfer, while different random seeds lead to high variance in target languages, the source language has much smaller variance in comparison even with standard Adam (Wu and Dredze, 2020).

Beyond optimizers, Mosbach et al. (2021) attribute high variance to generalization issues: despite having similar training loss, different models exhibit vastly different development set performance. However, in zero-shot cross-lingual transfer, the development or test performance variance is much smaller on the source language compared to target language.

3 Under-specified Optimization

Existing hypotheses do not explain the high variance of zero-shot cross-lingual transfer: much higher variance on generalization error of the target language compared to the source language. We propose a new explanation: *zero-shot cross-lingual transfer is an under-specified optimization problem.* Optimizing a multilingual model for a specific task using only source language annotation can choose from many good solutions. However, unbeknownst to the optimizer, these solutions have wildly different performance on the target language. Without the guidance of target data, the optimizer selects a solution that works for the source language without regard to its performance on target language test data. While we sometimes get lucky and the optimizer picks a solution good for both languages, many times the optimizer picks a solution that does poorly on the target language.

3.1 Linear Interpolation

We test this hypothesis via a linear interpolation between two models to explore the neural network parameter space. Consider three sets of neural network parameters: θ_{src}, θ_{tgt}, $\theta_{\{src,tgt\}}$ for a model trained on task data for the source language only, target language only and both languages, respectively. This includes both task-specific layers and encoders.[1] Note all three models have the same initialization before fine-tuning. We obtain the 1-dimensional (1D) linear interpolation of a monolingual (source) task trained model and bilingual task trained model with

$$\theta(\alpha) = \alpha\theta_{\{src,tgt\}} + (1-\alpha)\theta_{src} \quad (1)$$

or we could swap source and target by

$$\theta(\alpha) = \alpha\theta_{\{src,tgt\}} + (1-\alpha)\theta_{tgt} \quad (2)$$

where α is a scalar mixing coefficient (Goodfellow et al., 2014). Additionally, we can compute a 2-dimensional linear interpolation as

$$\theta(\alpha_1, \alpha_2) = \theta_{\{src,tgt\}} + \alpha_1\delta_{src} + \alpha_2\delta_{tgt} \quad (3)$$

where $\delta_{src} = \theta_{src} - \theta_{\{src,tgt\}}$, $\delta_{tgt} = \theta_{tgt} - \theta_{\{src,tgt\}}$, α_1 and α_2 are scalar mixing coefficients (Li et al., 2018).[2] Finally, we can evaluate any interpolated models on the development set of source and target languages, testing the generalization error on the same language and across languages.

The performance of the interpolated model illuminates the behavior of the model's parameters. Take Eq. (1) as an example: if the linear interpolated model performs consistently high for our task on the source language, it suggests that both models lie within the same local minimum of source language generalization error surface. Additionally,

[1]We experiment with interpolating the encoder parameters only and observe similar findings. On the other hand, interpolating the task-specific layer only has a negligible effect.

[2]Li et al. (2018) use two random directions and they normalize it to compensate scaling issue. In this setup, we find δ_{src} and δ_{tgt} have near identical norms, so we do not apply additional normalization. As these two directions are not random, we find that it spans around 55°. We plot the norm ratio and angle of these two vectors in App. B.

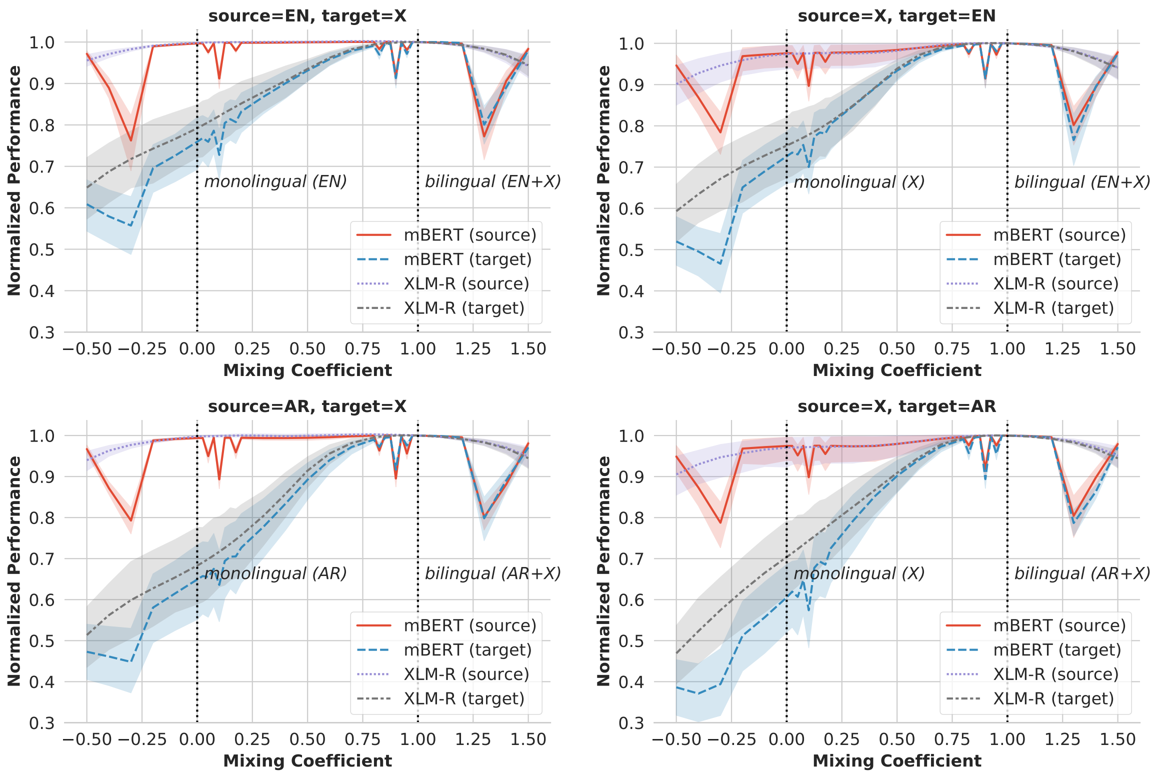


Figure 1: Normalized performance of a linear interpolated model between a monolingual and bilingual model. A single plot line shows the performance normalized by the matching bilingual model and aggregated over eight language pairs and four tasks, with the shaded region represents 95% confidence interval. The x-axis is the linear mixing coefficient α in Eq. (1) and Eq. (2), with $\alpha = 0$ and $\alpha = 1$ representing source language monolingual model and source + target bilingual model, respectively. Each subfigure title indicates the source and target languages. Across all experiments, the source language dev performance stay consistently high (red and purple lines) during interpolation while the target language dev performance starts low and increases smoothly as it moves towards the bilingual model (gray and blue lines). App. D breakdown this figure by tasks.

if the linear interpolated model performs vastly differently on the target language, it would support our hypothesis. On the other hand, if the linear interpolated model performance drops on the source language, it suggests that the local minimum of generalization error surface of monolingual model and bilingual model is linearly disconnected.

4 Experiments

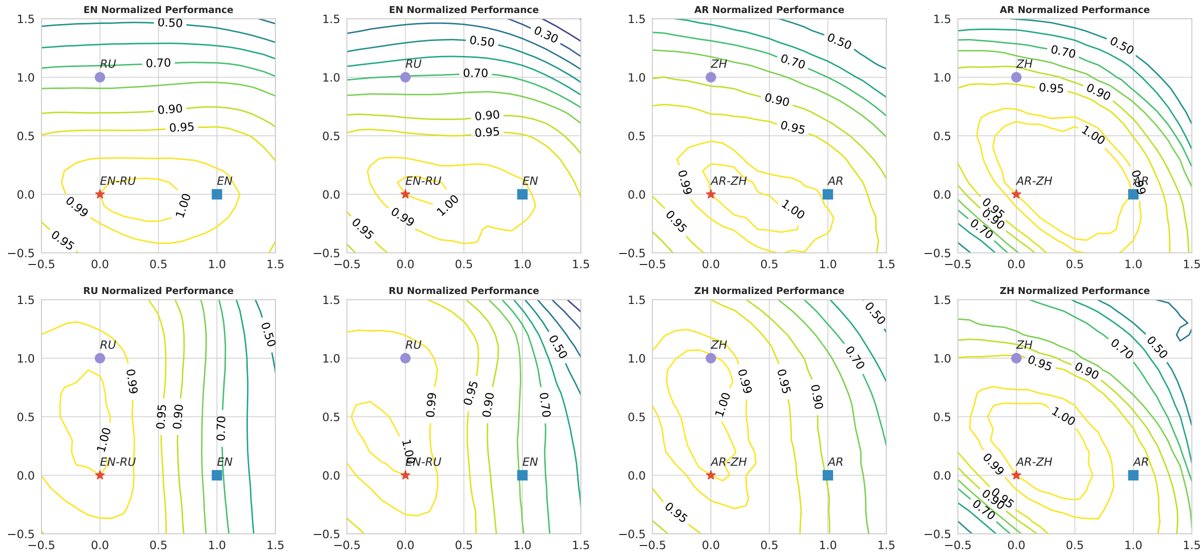
We consider four tasks: natural language inference (XNLI; Conneau et al., 2018), named entity recognition (NER; Pan et al., 2017), POS tagging and dependency parsing (Zeman et al., 2020). We evaluate XNLI and POS tagging with accuracy (ACC), NER with span-level F1, and parsing with labeled attachment score (LAS). We consider two encoders: base mBERT and large XLM-R. For the task-specific layer, we use a linear classifier for XNLI, NER, and POS tagging, and Dozat and Manning (2017) for dependency parsing.

To avoid English-centric experiments, we consider two source languages: English and Arabic. We choose 8 topologically diverse target languages: Arabic[3], German, Spanish, French, Hindi, Russian, Vietnamese, and Chinese. We train the source language only and target language only monolingual model as well as a source-target bilingual model.

We compute the linear interpolated models as described in §3.1 and test it on both the source and target language development set. We loop over $\{-0.5, -0.4, \cdots, 1.5\}$ for α, α_1 and α_2.[4] We report the mean and variance of three runs by using different random seeds. We normalized both mean and variance of each interpolated model by the bilingual model performance, allowing us to aggregate across tasks and language pairs. Details of fine-tuning can be found in App. A.

[3] Arabic is only used when English is the source language.

[4] We additionally select 0.025, 0.05, 0.075, 0.125, 0.15, 0.175, 0.825, 0.85, 0.875, 0.925, 0.95, and 0.975 for α due to preliminary experiment.



(a) EN-RU NER w/ mBERT (b) EN-RU NER w/ XLM-R (c) AR-ZH XNLI w/ mBERT (d) AR-ZH XNLI w/ XLM-R

Figure 2: Normalized performance of 2D linear interpolation between bilingual model and monolingual models. The x-axis and the y-axis are the α_1 and α_2 in Eq. (3), respectively. By comparing mBERT and XLM-R, we observe that XLM-R has flatter target language generalization error surface compared to mBERT. Different language pairs and tasks combination shows similar trends and additional figures can be found in App. E

5 Results

In Fig. 1, we observe that interpolations between the source monolingual and bilingual model have consistently similar source language performance. In contrast, the target language performance smoothly improves as the interpolated model moves from the zero-shot model to bilingual model.[5] The only exception is mBERT, where the performance drops slightly around 0.1 and 0.9 locally. This contrast, XLM-R has a flatter slope and smoother interpolated models. It suggests that the source monolingual model and bilingual model are linearly connected on the source language generalization error surface, and any model in this local minimum performs equally well on the source language. However, the target language performance differs significantly. Due to high solution space dimensionality, training with source alone is unlikely to find this smaller subset of solutions by chance.

Fig. 2 further demonstrates this finding with a 2D linear interpolation. The generalization error surface of the target language of XLM-R is much flatter compared to mBERT, perhaps the fundamental reason why XLM-R performs better than mBERT in zero-shot transfer, similar to findings in other computer vision models (Li et al., 2018). As we discuss in §3, these two findings support our hypothesis that zero-shot cross-lingual transfer is an under-specified optimization problem.

6 Discussion

We have presented evidence that zero-shot cross-lingual transfer is an under-specified optimization problem, and the cause of high variance on target language but not the source language tasks during zero-shot cross-lingual transfer. This finding holds across 4 tasks, 2 source languages and 8 target languages. Training bigger encoders addresses this issue indirectly by producing encoders with flatter cross-lingual generalization error surfaces. However, a more robust solution may be found in the future by introducing constraints into the optimization problem that directly addresses the under-specification of the optimization.

There are a few potential solutions. Few-shot cross-lingual transfer (Zhao et al., 2021) or silver target data (Yarmohammadi et al., 2021) can provide useful constraints. Unsupervised model selection (Chen and Ritter, 2020) and optimization regularization (Aghajanyan et al., 2021) add constraints without annotation. Perhaps a combination of the above techniques might complement each other and further constrain the optimization problem.

[5] We also show the variance of the interpolated models in App. C. The source language has much lower variance compared to target language on the monolingual side of the interpolated models, echoing findings in Wu and Dredze (2020).

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Yang Chen and Alan Ritter. 2020. Model selection for cross-lingual transfer using a learned scoring function. *arXiv preprint arXiv:2010.06127*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. 2014. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Online. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, et al. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. *arXiv preprint arXiv:2109.06798*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample BERT fine-tuning. In *International Conference on Learning Representations*.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A Fine-tuning Experiments Detail

We follow the implementation and hyperparameter of Wu and Dredze (2020). We optimize with Adam (Kingma and Ba, 2014). The learning rate is `2e-5`. The learning rate scheduler has 10% steps linear warmup then linear decay till 0. We train for 5 epochs and the batch size is 32. For token level tasks, the task-specific layer takes the representation of the first subword, following previous work (Devlin et al., 2019; Wu and Dredze, 2019). Model selection is done on the corresponding dev set of the training set.

During fine-tuning, the maximum sequence length is 128. We use a sliding window of context to include subwords beyond the first 128 for NER and POS tagging. At test time, we use the same maximum sequence length with the exception of parsing, where the first 128 words instead of subwords of a sentence were used. We ignore words with POS tags of `SYM` and `PUNCT` during parsing evaluation. For NER, the prediction of `BIO` was post-processed to make sure a valid span is produced.

All datasets we used are publicly available: NER[6], XNLI[7,8], POS tagging and dependency parsing[9]. For POS tagging and dependency parsing, we use the following treebanks: Arabic-PADT, German-GSD, English-EWT, Spanish-GSD, French-GSD, Hindi-HDTB, Russian-GSD, Vietnamese-VTB, and Chinese-GSD. Data statistic can be found in Tab. 1.

	XNLI	NER	POS tagging Parsing
en-train	392703	20000	12543
en-dev	2490	10000	2002
ar-train	392703	20000	6075
ar-dev	2490	10000	909
de-train	392703	20000	13814
de-dev	2490	10000	799
es-train	392703	20000	14187
es-dev	2490	10000	1400
fr-train	392703	20000	14449
fr-dev	2490	10000	1476
hi-train	392703	5000	13304
hi-dev	2490	1000	1659
ru-train	392703	20000	3850
ru-dev	2490	10000	579
vi-train	392703	20000	1400
vi-dev	2490	10000	800
zh-train	392703	20000	3997
zh-dev	2490	10000	500

Table 1: Number of examples.

B Norm Ratio and Angle of δ_{src} and δ_{tgt}

Fig. 3 plots the relationship between $\|\delta_{src}\|/\|\delta_{tgt}\|$ and angle between δ_{src} and δ_{tgt}. We observe most δ_{src} and δ_{tgt} have similar norms, and the angle between them is around 55°.

C Normalized Variance of Linear Interpolated Models

Fig. 4 plots the normalized variance of linear interpolated models. We observe that the source language has much lower variance compared to target language on the monolingual side of the interpolated models

D Breakdown of Normalized Performance of Linear Interpolated Models by Tasks

Fig. 5 (NER), Fig. 6 (Parsing), Fig. 7 (POS), and Fig. 8 (XNLI) plot the normalized performance of linear interpolated models breakdown by task. We observe similar findings as Fig. 1.

E Additional 2D Linear Interpolation

Fig. 9 plots additional 2D linear interpolation. We observe similar findings as Fig. 2.

[6] https://www.amazon.com/clouddrive/share/d3KGCRCIYwhKJF0H3eWA26hjg2ZCRhjpEQtDL70FSBN

[7] https://dl.fbaipublicfiles.com/XNLI/XNLI-MT-1.0.zip

[8] https://dl.fbaipublicfiles.com/XNLI/XNLI-1.0.zip

[9] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3424

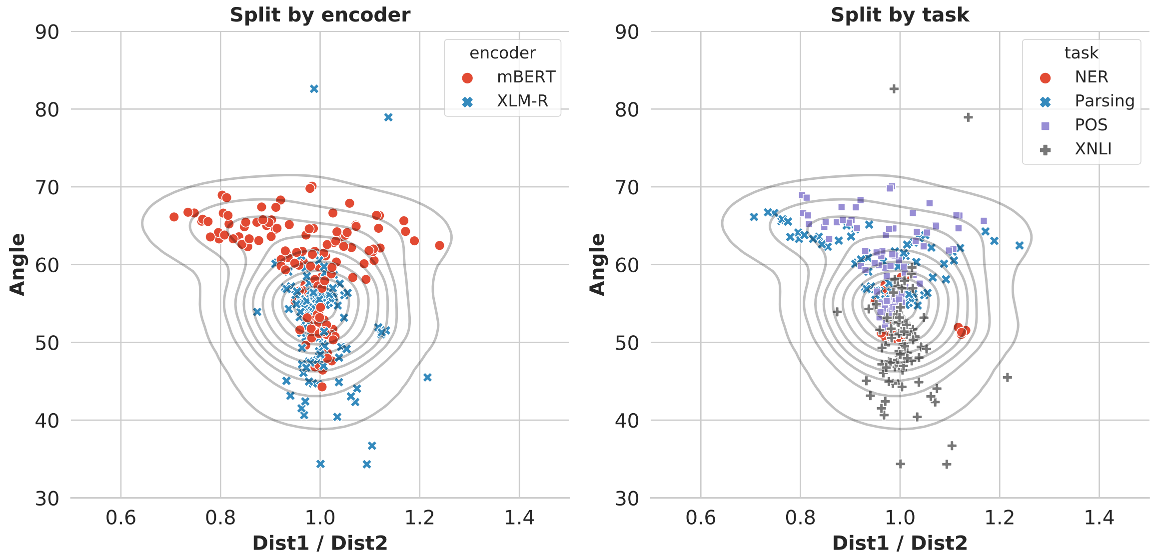


Figure 3: $\|\delta_{src}\|/\|\delta_{tgt}\|$ v.s. angle between δ_{src} and δ_{tgt}. Most δ_{src} and δ_{tgt} have similar norms, and the angle between them is around 55°.

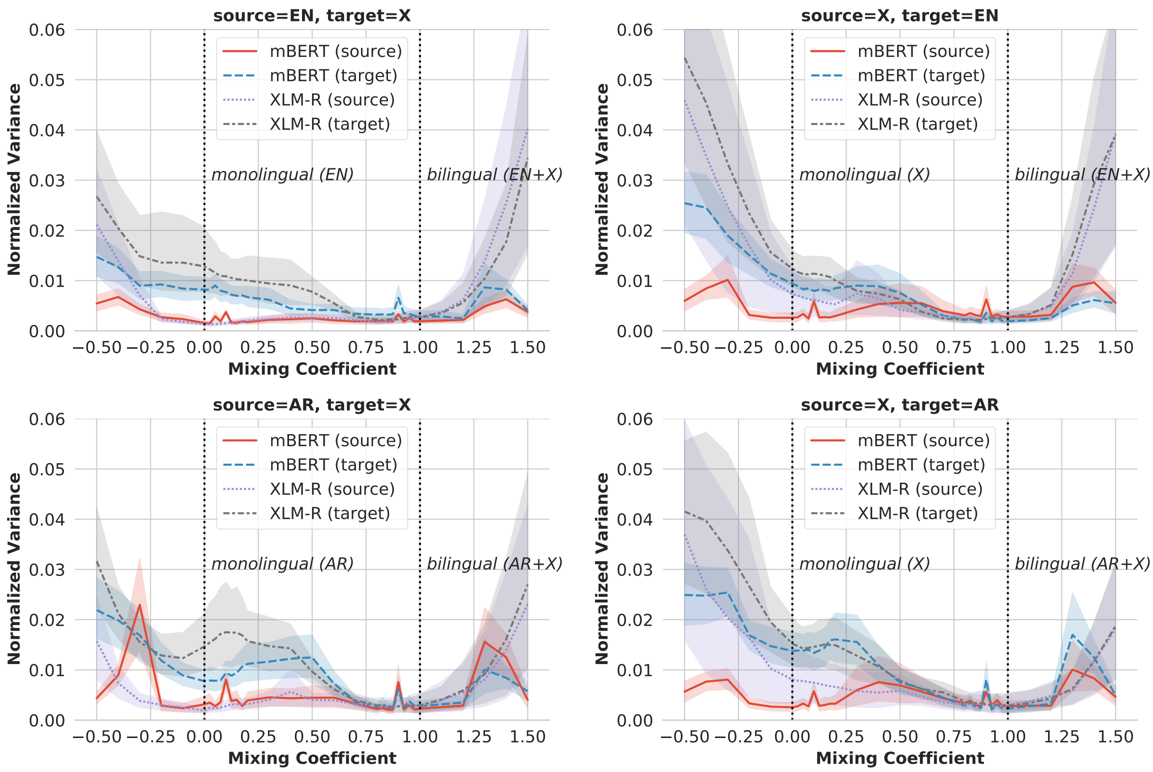


Figure 4: Normalized variance of linear interpolation between monolingual model and bilingual model. The source language has much lower variance compared to target language on the monolingual side of the interpolated models.

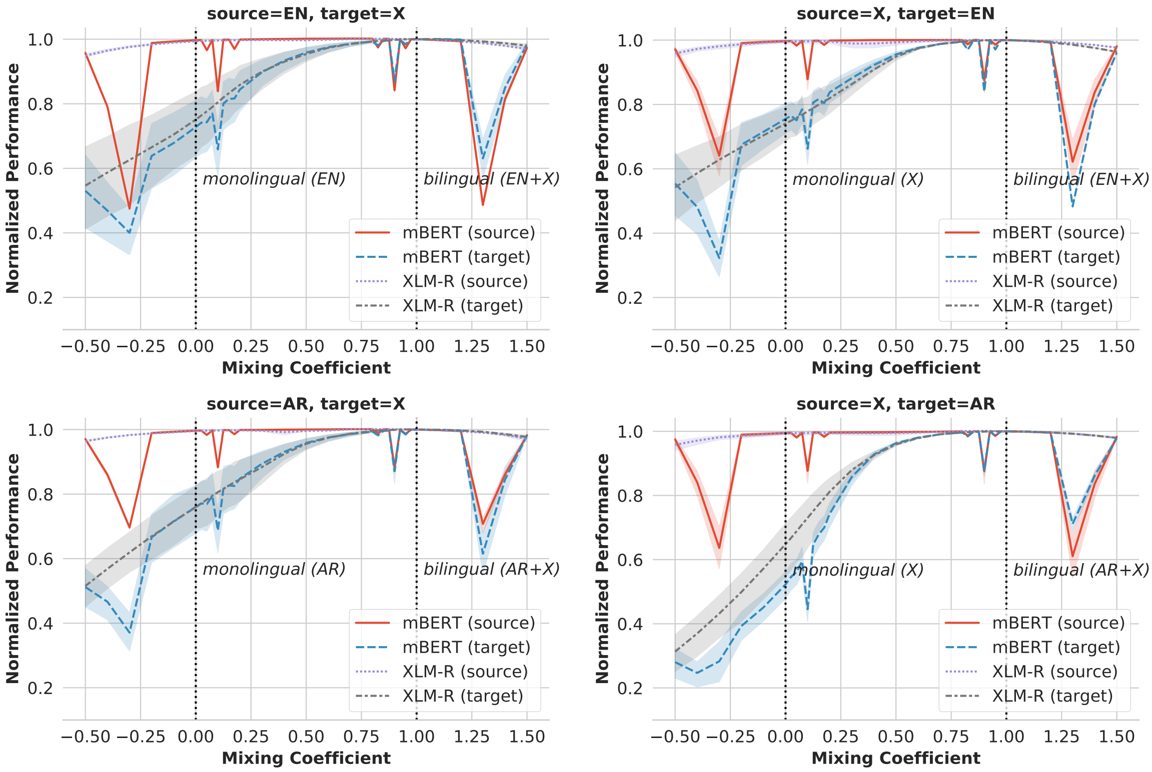


Figure 5: Normalized NER performance of linear interpolated model between monolingual and bilingual model

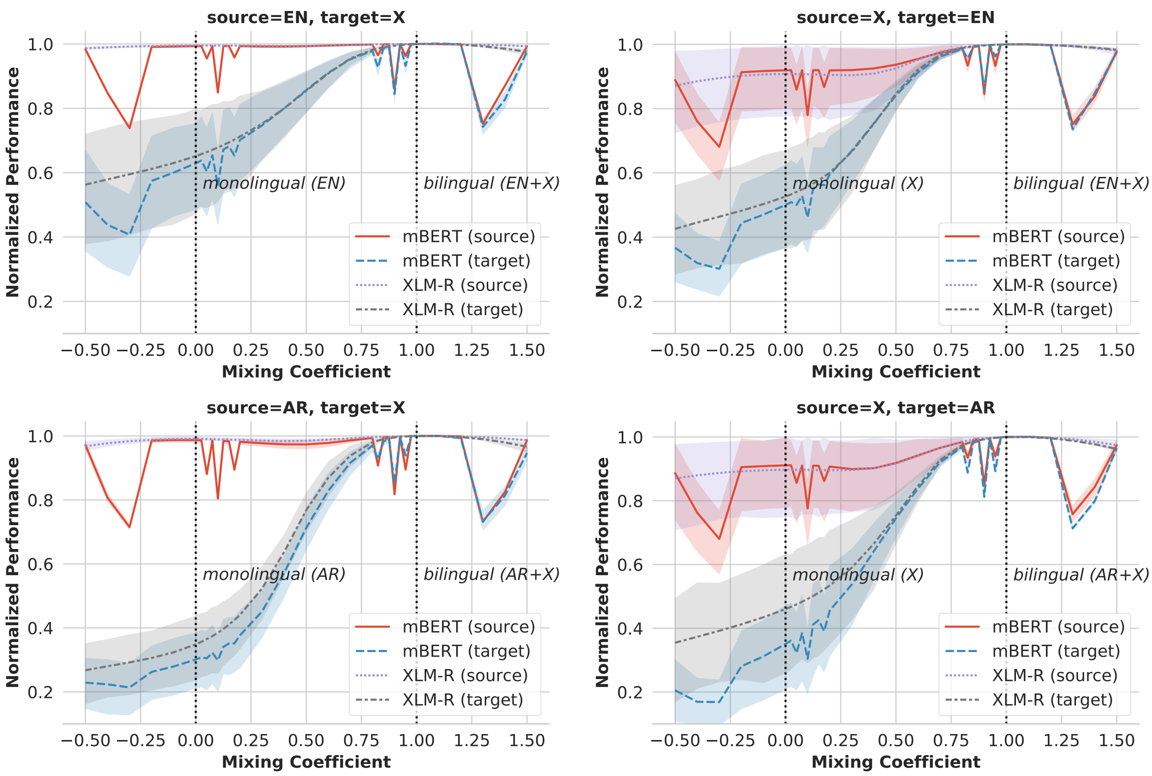


Figure 6: Normalized Parsing performance of linear interpolated model between monolingual and bilingual model

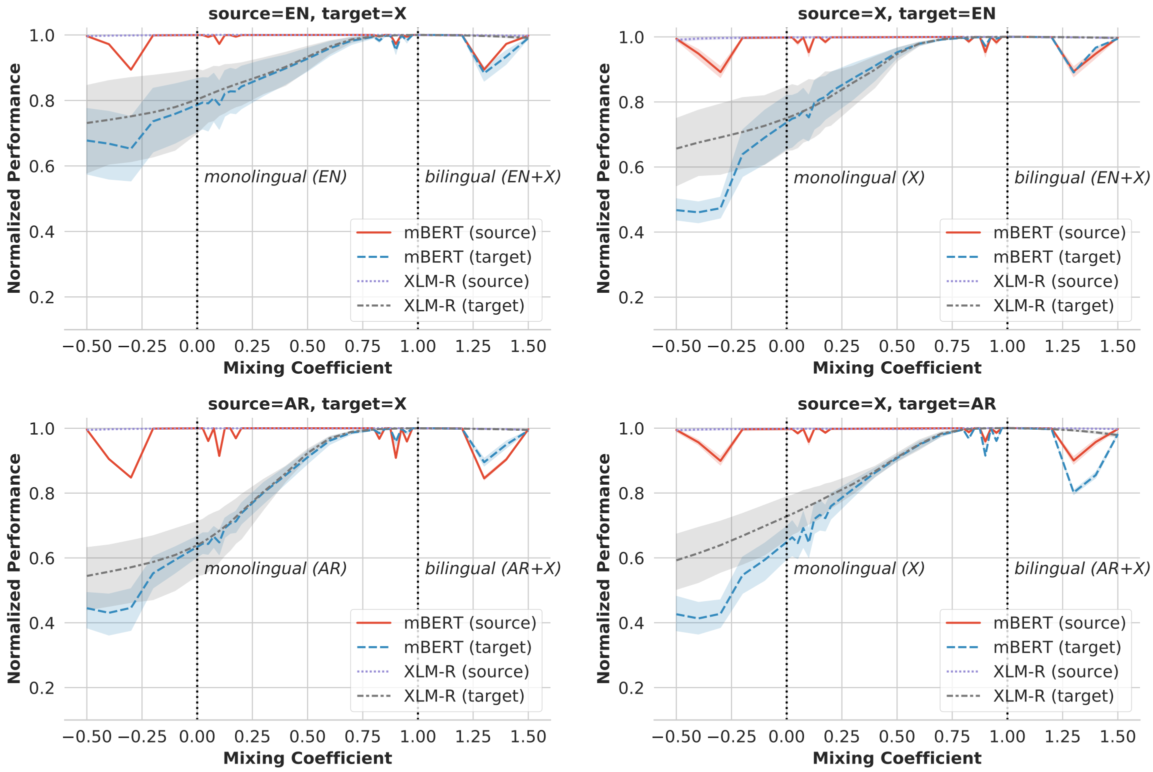


Figure 7: Normalized POS performance of linear interpolated model between monolingual and bilingual model

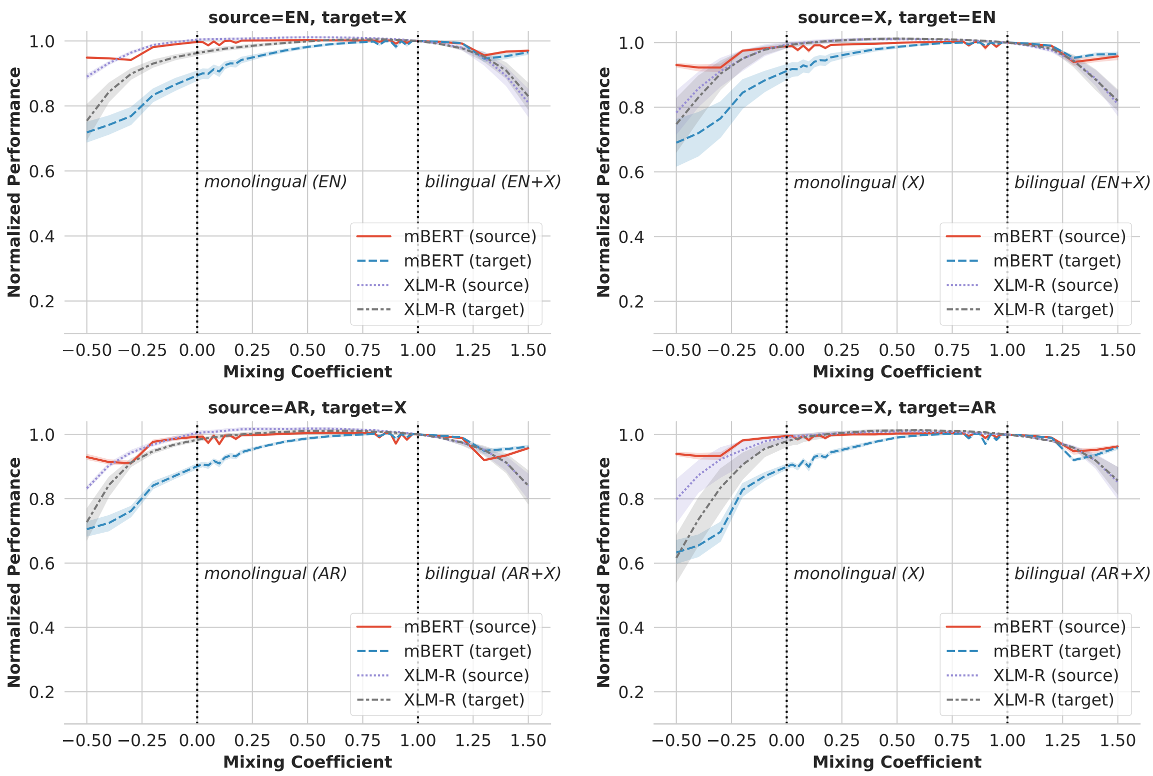
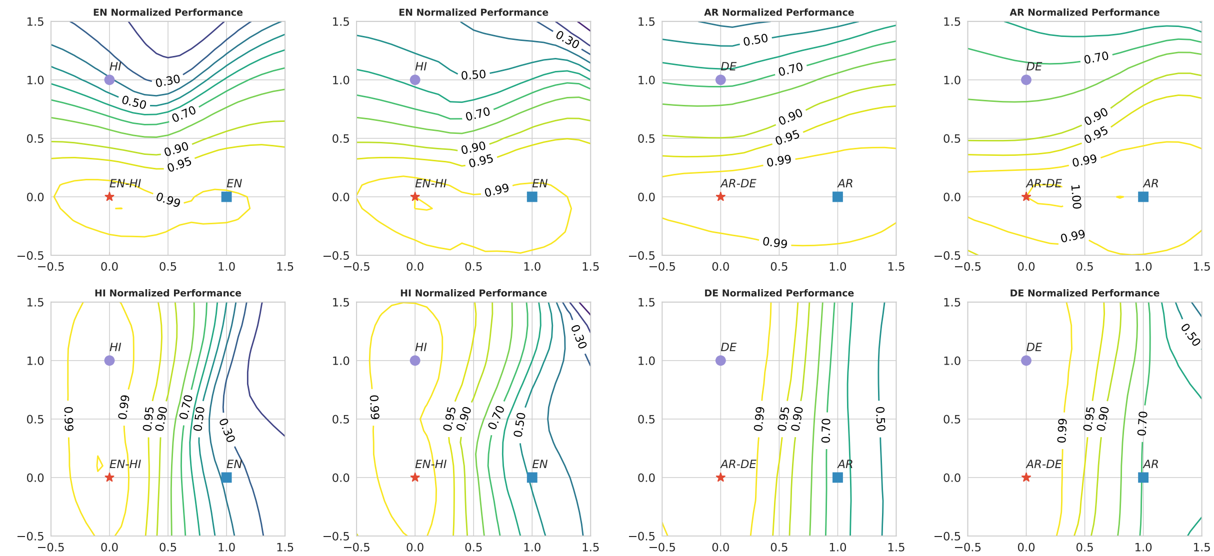


Figure 8: Normalized XNLI performance of linear interpolated model between monolingual and bilingual model



(a) EN-HI Parsing w/ mBERT (b) EN-HI Parsing w/ XLM-R (c) AR-DE POS w/ mBERT (d) AR-DE POS w/ XLM-R

Figure 9: Additional normalized performance of 2D linear interpolation between bilingual model and monolingual models