Analisando o Fine-Tuning para Transferência de Conhecimento entre Línguas: um estudo para a Língua Portuguesa

Yuri Hughes [1] **, Marlo Souza** [1]

[1]Instituto de Computação – Universidade Federal da Bahia (UFBA)

Resumo. *Grandes Modelos de Linguagem (LLMs) desempenham um papel central no cenário moderno de PLN, devido aos seus consistentes bons resultados em muitas tarefas na literatura. No entanto, o treinamento desses modelos envolve altos custos associados. Devido à escassez de dados necessários para tarefas em idiomas com poucos recursos, a literatura tem proposto o uso de tecnologia multilíngue combinada com técnicas de transferência de conhecimento entre línguas. Esta pesquisa explora o fine-tuning de modelos para transferência de conhecimento entre línguas para o português, em diversas tarefas, avaliando os trade-offs entre quantidade de dados, recursos computacionais, tempo de treinamento e desempenho do modelo.*

1. Introdução

Grandes Modelos de Linguagem (LLMs) se destacam em diversas aplicações de Processamento de Linguagem Natural devido à fácil adaptação a diversas tarefas por meio de fine-tuning. No entanto, seu pré-treinamento exige uma quantidade significativa de dados linguísticos, o que torna desafiador o desenvolvimento de tecnologia para idiomas sub-representados e com recursos limitados.

Adicionalmente, os custos de pré-treinamento de tais modelos são bastante elevados. Por exemplo, o BERTimbau [Souza et al. 2020], um modelo monolíngue de referência para o português, levou 4 dias para ser treinado em sua versão base e 7 dias para a versão grande, usando uma TPU v3-8, sobre o conjunto de dados Brazilian Web Corpus (BrWaC) [Zilio and Rino 2011]. Esse nível de poder computacional frequentemente está fora do alcance de muitos pesquisadores, especialmente aqueles pertencentes a comunidades marginalizadas.

Uma possível solução, como sugerido por pesquisas recentes, é aproveitar as capacidades de transferência de conhecimento das LLMs para a transferência entre línguas. No entanto, estudos recentes indicam que, embora os modelos multilíngues consigam lidar com tarefas de PLN em diversas línguas, eles geralmente apresentam desempenho inferior em comparação com seus equivalentes monolíngues [Wu and Dredze 2019, Martin et al. 2020, Souza et al. 2020].

Este estudo investiga a transferência de conhecimento entre línguas em LLMs através da sua avaliação em diversas tarefas em português, incluindo Reconhecimento de Entidades Nomeadas, Análise de Sentimentos e Inferência em Linguagem Natural. Para tanto, empregamos modelos treinados com dados em inglês para uma dada tarefa, à medida que realizamos o *fine-tuning* sobre os dados em português para a tarefa em questão, considerando diferentes cenários de disponibilidade de dados e analisando as medidas de desempenho obtidas. Além disso, acompanha-se o uso geral de recursos, como volume de dados, custos computacionais e tempo necessário para alcançar os resultados.

2. Trabalhos Relacionados

O aprendizado por transferência no Processamento de Linguagem Natural (PLN) foi inspirado em seu sucesso na área de Visão Computacional. Tais modelos, como o ULMFiT

[Howard and Ruder 2018], propõem a ideia de um modelo geral adaptado para aprender estruturas linguísticas latentes nos dados que podem ser posteriormente empregados para outras tarefas. O ULMFiT estabeleceu um processo em duas etapas: pré-treinamento em domínio geral com grandes corpora de texto, seguido de fine-tuning específico para a tarefa. Essa abordagem serviu de base para modelos posteriores, incluindo o BERT [Devlin et al. 2019].

Embora alguns trabalhos tenham investigado a capacidade de modelos multilíngues, como o mBERT [Devlin et al. 2019], para transferência de conhecimento entre línguas, como no estudo de [Wu and Dredze 2019], estudos mais recentes mostram que modelos monolíngues frequentemente superam os modelos multilíngues em diversas tarefas [Martin et al. 2020, Souza et al. 2020]. Razuvayevskaya et al. [Razuvayevskaya et al. 2024] apresentam uma avaliação de transferência de conhecimento nesses modelos considerando técnicas como fine-tuning, adaptadores [Pfeiffer et al. 2020] e LoRA [Hu et al. 2021] em conjuntos multilíngues e entre línguas, no contexto de classificação de artigos jornalísticos. O trabalho relata que o fine-tuning supera as predições zero-shot entre línguas em muitos experimentos, embora o LoRA se destaque em cenários com dados limitados.

Neste trabalho, investigamos a eficiência da abordagem de fine-tuning para transferência de conhecimento entre línguas, considerando diferentes cenários de disponibilidade de dados e diferentes tarefas na área de PLN para língua portuguesa, levando em conta seus custos computacionais associados.

3. Métodos e Experimentos

Nesta seção, apresentamos o fluxo de trabalho adotado para o treinamento e a avaliação dos nossos modelos, descrevendo os conjuntos de dados utilizados, tanto os de origem quanto os de destino.

3.1. Seleção do Modelo

Nossa abordagem utiliza o modelo multilíngue mBERT-base [Devlin et al. 2019], um codificador bidirecional baseado em Transformer, pré-treinado em dados da Wikipédia de 104 idiomas, incluindo português e inglês. Selecionamos este modelo devido ao seu bom desempenho em pesquisas acadêmicas sobre tarefas multilíngues, aliado a um custo computacional menor em comparação com modelos maiores. Além disso, o mBERT possui um vocabulário compartilhado baseado em word-pieces que engloba representações de subpalavras entre diferentes línguas, o que é especialmente vantajoso para tarefas entre idiomas.

3.2. Datasets

Utilizamos cinco datasets no total, para três tarefas de PLN: Reconhecimento de Entidades Nomeadas (NER), usando o dataset Universal NER[1]; Inferência de Linguagem Natural (NLI), usando o dataset Multilingual NLI 27l [2]; e Análise de Sentimentos (SA), com base nos datasets IMDB[3] e IMDB-PT[4]. A escolha das tarefas se deve à variedade dos tipos de tarefa (Classificação de Texto e Rotulagem Sequencial), à disponibilidade de corpora com anotações compatíveis para dados em inglês e português, além da relevância das tarefas.

[1] https://huggingface.co/datasets/universalner/universal_ner
[2] https://huggingface.co/datasets/MoritzLaurer/multilingual-NLI-26lang-2mil7
[3] https://huggingface.co/datasets/mteb/imdb
[4] https://huggingface.co/datasets/celsowm/imdb-reviews-pt-br

Tabela 1. Datasets utilizados e número de exemplos de treinamento

Tarefa	Língua	Dataset	# Número de instâncias
NER	Inglês	Universal NER - EWT	12543
NER	Português	Universal NER - Bosque	7018
SA	Inglês	IMDB	25000
SA	Português	IMDB-PT	49459 (39567)
NLI	Inglês e Português	Multilingual NLI 27l	25000

Na Tabela 1, observe que, como o conjunto de dados IMDB-PT não possui uma partição de teste, 80% do dataset foi utilizado para o treinamento dos modelos no experimento (número de exemplos indicado entre parênteses), enquanto 20% foi reservado para avaliação.

3.3. Métricas

Os modelos foram avaliados com base em duas métricas: medida F1 e tempo de treinamento. Para a tarefa de NER, o F1 foi calculado tomando a avaliação geral de todas as classes segundo a metodologia SeqEval [Ramshaw and Marcus 1995] e foi computado utilizando a biblioteca evaluate da HuggingFace.

4. Experimentos e Resultados

Esta seção detalha a metodologia experimental e a implementação do fine-tuning. Os experimentos foram realizados em duas máquinas: os experimentos relacionados às tarefas de NER e NLI foram conduzidos em uma máquina com CPU Intel i7 de 13ª geração, 64 GB de RAM e GPU NVidia RTX 3090 de 24 GB; já os experimentos de SA foram realizados em uma máquina com CPU Intel(R) Xeon(R) Gold 6346, 93 GB de RAM e GPU NVidia A4000 de 16 GB.

4.1. Configuração Experimental

Nosso esquema, comum a todos os experimentos, segue as seguintes configurações:

- Carregamento e Pré-processamento dos Dados: os conjuntos de dados, tokenizadores e modelos de linguagem foram obtidos através das bibliotecas transformers e datasets da HuggingFace.
- Seleção de Hiperparâmetros experimentais: os hiperparâmetros de treino, a saber, número de épocas, tamanho de batch, taxa de aprendizado, decaimento de peso e *warm-up ratio*, foram selecionados pela otimização de desempenho no treinamento do modelo sobre os dados em inglês. Para tanto, utilizamos o framework Optuna [Akiba et al. 2019], com 10 execuções, maximizando a métrica de F1 nos conjuntos de treino e teste em inglês. Os melhores hiperparâmetros encontrados foram mantidos para o restante do processo de treinamento sobre os dados em português.
- Treinamento sobre os dados em português: o melhor modelo previamente ajustado à tarefa-alvo em inglês, i.e. que obteve melhor métrica F1 no ajuste de hiperparâmetros, foi posteriormente treinado por fine-tunning utilizando apenas os dados em português, considerando diferentes cenários de disponibilidade de dados na língua-alvo. Para tanto, variamos a quantidade de dados do conjunto de treino apresentados ao modelo, selecionando uma partição contendo uma proporção variável dos dados de treinamento em português (0%, 10%, 30%, 50%, 70% e 90%). Para cada proporção, foram realizadas cinco execuções independentes do processo de fine-tunning, a fim de avaliar possíveis variações nos resultados de avaliação.

Tabela 2. Configuração de hiperparâmetros para o experimento de fine-tuning

Parâmetro	SA	NER	NLI
Tamanho de Batch	32	64	64
número de épocas	3	7	7
Warmup	0.1922	0.2187	0.1463
Decaimento	0.0444	0.0275	0.0167
taxa de aprendizado	3.3227e-5	4.6166e-5	7.6113e-5

4.2. Resultados

A seguir, apresentamos os resultados de nossos experimentos, reportando a métrica F1 e o Tempo de Treinamento por época (TT, em segundos), com os desvios padrão indicados entre parênteses. Os desvios padrão foram omitidos quando o dígito mais significativo encontra-se além da segunda casa decimal.

Tabela 3. Resultados do fine-tuning nas tarefas (SA, NER, NLI)

% Data	SA		NER		NLI	
	F1	TT	F1	TT	F1	TT
0%	0.85	0.00	0.78	0.00	0.75	0.00
10%	0.90	154.13 (0.08)	0.86	14.35 (0.06)	0.73	40.36 (0.01)
30%	0.92	462.1 (0.2)	0.86	43.16 (0.06)	0.74	121.96 (0.05)
50%	0.92	770.5 (0.5)	0.88	72.15 (0.10)	0.75	203.4 (0.1)
70%	0.93	1078.6 (0.7)	0.88	101.11 (0.04)	0.74	285.1 (0.3)
90%	0.93	1385 (2)	0.89	130.16 (0.09)	0.75	367.0 (0.2)

5. Conclusões

Este trabalho investigou a eficácia do fine-tuning para Transferência de Conhecimento entre Línguas (Cross-language Knowledge Transfer) para a língua portuguesa em diferentes tarefas. Focamos em duas tarefas baseadas em classificação de texto - uma fortemente associada a informações lexicais (Análise de Sentimentos) e outra relacionada à semântica em nível de sentença (Inferência em Linguagem Natural); além de uma tarefa baseada em rotulagem de sequência (Reconhecimento de Entidades Nomeadas), na qual o modelo depende tanto de pistas lexicais quanto estruturais.

Nossos resultados mostram que é possível obter um desempenho competitivo com uma quantidade substancialmente reduzida de dados na língua-alvo, em comparação com o treinamento monolíngue completo. Essa abordagem reduz significativamente os custos computacionais, ao mesmo tempo em que mantém a qualidade das predições.

Para trabalhos futuros, nosso objetivo é explorar essa abordagem no treinamento e avaliação de outras arquiteturas de modelagem de linguagem, seja baseadas em Transformers ou em atenção linear, e em outras tarefas e paradigmas. Este estudo nos ajudará a compreender quais são os limites e as aplicações reais dos modelos de linguagem multilíngues em tarefas específicas.

6. Agradecimentos

Agradecemos ao Programa Institucional de Bolsas de Iniciação Científica da UFBA (PIBIC-UFBA) pela bolsa que viabilizou o desenvolvimento do trabalho, à Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) por auxílio financeiro através do projeto TIC 002/2015 no qual tal pesquisa se insere e à CAPES através do auxílio financeiro 001 para Programas de Pós-Graduação.

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *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.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, W., and Smola, A. (2021). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., De La Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Razuvayevskaya, O., Wu, B., Leite, J. A., et al. (2024). Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *PLOS ONE*, 19(5):e0301738.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pre-trained bert models for brazilian portuguese. *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 833–844. Association for Computational Linguistics.

Zilio, L. V. and Rino, L. M. (2011). Brwac: A new brazilian web corpus. *PROPOR*, pages 149–158.