Aplicación de Reconocimiento de Formas

Traducción Automática Mültilingue mediante LLMs

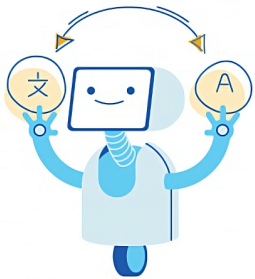
Jorge Iranzo Sánchez

Valencia,
15 de Febrero 2023

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Índice

Introducción



- *Boom* de modelos de lenguajes de tamaño masivo basados en Transformers [Zha+23] [Bro+20].
- Eficaces en una multitud de tareas.
- Aunque no se hayan entrenado explícitamente para esta, ¿funcionan para la Traducción Automática?

Trabajo relacionado

- MT como tarea de modelado lingüístico → Modelos entrenados concatenando frases de origen y de destino en *train* [Wan+21] [Gao+22].
- *Scaling laws* en MT: [GDK21] [Gho+21] [Gho+22].
- Multiples estudios para LLMs: General [Zhu+23], GPTs [Hen+23] [GWH23] [MHW23], BLOOM [BY23], GLM [ZHB23], PALM [Gar+23] ...
- Se centran en explorar aspectos sobre *prompts* o multilingualidad.

Tarea

- Evaluación de LLMs respecto *baseline* de Transformers BIG [Vas+17] y NLLB [Cos+22].
- Dirección: en $\rightarrow$ {fr, sl}.
- Objetivo: Traducción de charlas biomédicas, Proyecto INTERACT.

Arquitectura

- Baseline: Vocabulario mediante SentenciePiece.
- NLLB: Tranformer Enc-Decs Multilingües en 200+ idiomas.
  - Escalado masivo en la recopilación de datos y limpieza.
  - Destaca el uso en el modelo mas grande de Mixtura de Expertos (MoE) [Sha+17].
- BLOOM [Sca+22]: Transformer *decoder* multilingüe entrenado en un corpus masivo de 1.6TB, ROOTS [Lau+23].
  - Elegimos la versión del modelo con 3B params, quantizado a INT-8.
  - ROOTS **no** contiene datos en esloveno.

Datos

- Baseline: Entrenados sobre corpus de OPUS [Tie12] en Fairseq [Ott+19]. fr $\rightarrow$ 256M, sl $\rightarrow$ 256M,
- Dev/Test: $\sim$ 1.4K frases.
- Dataset **aux** de 430k frases limpiado mediante Bicleaner y Bifixor [Ram+20].

Diseño experimental

- Fine-tuning mediante LORA [Hu+22] en Europarl-ST [Ira+20].
- Evaluación con BLEU y COMET-22 [Rei+22]:
  - *n*-gramas *vs* neuronal.
  - *Findings* de WMT22: **Stop Using BLEU** [Fre+22].
- Prompt [BY23]:
  - *Translate from [src] to [tgt]:\n [src] [src_sentence] = [src_tgt]:*
- Selección inteligente de *shots* para BLOOM: *k*-nn mediante FAISS [JDJ19] de embeddings con LaBSE [Fen+22] sobre **aux**.
- Beam search con $k = 5$.

Resultados

Modelo	Shots	LORA	Francés		Esloveno	
			BLEU	COMET-22	BLEU	COMET-22
Baseline-300M			51.5	81.8	40.8	84.9
NLLB-600M			53.5	82.1	33.3	83.1
Bloom-3B	0	✗	12.6	65.6	✗	
	1		36.8	78.9		
	2		38.0	77.0		
	3		39.4	79.0		
⟶	0	✓	45.5	82.1	✗	
	1		38.1	78.0		
	2		36.1	80.3		
	3		32.9	78.6		

Resultados (cont.)

			Francés		Esloveno	
Modelo	Shots	LORA	BLEU	COMET-22	BLEU	COMET-22
NLLB-1.3B			55.6	82.81	36.2	84.97
NLLB-3.3B			56.4	82.84	38.8	85.19
Bloom-175B	0	✗	45.3	80.98	✗	
	1		48.8	82.17		

Discusión y Conclusiones

- LLMs: Buenos resultados, pero con cierta inestabilidad.
- LORA es una manera eficaz de adaptar LLMs a MT, pero hay que tener cuidado con el formato del train.
- No todas las lenguas en el MT multilingüe son iguales.
- Cuanto mas grande y mas datos, mejor.

Trabajo Futuro

- Más pares de lenguas y modelos.
- ¿Y cuantizar con FP-4? [Det+23]
- ¿Y *promtps* basados en diccionarios? [GGZ23]
- LLMs no están entrenadas explícitamente para MT. Pero, ¿cuánto *bitext* han visto? [BCF23]

Gracias por su atención

Bibliografía I

[BCF23] Eleftheria Briakou, Colin Cherry y George F. Foster. “Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM’s Translation Capability”. En: *CoRR* abs/2305.10266 (2023).

[Bro+20] Tom B. Brown et al. “Language Models are Few-Shot Learners”. En: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. por Hugo Larochelle et al. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bibliografía II

[BY23] Rachel Bawden y Franccois Yvon. “Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM”. En: *ArXiv* abs/2303.01911 (2023).

[Cos+22] Marta R. Costa-jussà et al. “No Language Left Behind: Scaling Human-Centered Machine Translation”. En: *CoRR* abs/2207.04672 (2022).

[Det+23] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG].

[Fen+22] Fangxiaoyu Feng et al. “Language-agnostic BERT Sentence Embedding”. En: *ACL (1)*. Association for Computational Linguistics, 2022, págs. 878-891.

Bibliografía III

[Fre+22] Markus Freitag et al. “Results of WMT22 Metrics Shared Task: Stop Using BLEU - Neural Metrics Are Better and More Robust”. En: *WMT*. Association for Computational Linguistics, 2022, págs. 46-68.

[Gao+22] Yingbo Gao et al. “Is Encoder-Decoder Redundant for Neural Machine Translation?” En: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*. Ed. por Yulan He et al. Association for Computational Linguistics, 2022, págs. 562-574. URL: https://aclanthology.org/2022.aacl-main.43.

Bibliografía IV

[Gar+23] Xavier García et al. “The unreasonable effectiveness of few-shot learning for machine translation”. En: *ArXiv* abs/2302.01398 (2023).

[GDK21] Mitchell A. Gordon, Kevin Duh y Jared Kaplan. “Data and Parameter Scaling Laws for Neural Machine Translation”. En: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. por Marie-Francine Moens et al. Association for Computational Linguistics, 2021, págs. 5915-5922. DOI: 10.18653/v1/2021.emnlp-main.478. URL: https://doi.org/10.18653/v1/2021.emnlp-main.478.

Bibliografía V

[GGZ23] Marjan Ghazvininejad, Hila Gonen y Luke Zettlemoyer. “Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation”. En: *ArXiv* abs/2302.07856 (2023).

[Gho+21] Behrooz Ghorbani et al. *Scaling Laws for Neural Machine Translation*. 2021. arXiv: 2109.07740 [cs.LG].

[Gho+22] Behrooz Ghorbani et al. “Scaling Laws for Neural Machine Translation”. En: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL: https://openreview.net/forum?id=hR%5C_SMu8cxCV.

[GWH23] Yuan Gao, Ruili Wang y Feng Hou. “Unleashing the Power of ChatGPT for Translation: An Empirical Study”. En: *ArXiv* abs/2304.02182 (2023).

Bibliografía VI

[Hen+23] Amr Hendy et al. “How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation”. En: *ArXiv* abs/2302.09210 (2023).

[Hu+22] Edward J. Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. En: *ICLR*. OpenReview.net, 2022.

[Ira+20] Javier Iranzo-Sánchez et al. “Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates”. En: *ICASSP*. IEEE, 2020, págs. 8229-8233.

[JDJ19] Jeff Johnson, Matthijs Douze y Hervé Jégou. “Billion-scale similarity search with GPUs”. En: *IEEE Transactions on Big Data* 7.3 (2019), págs. 535-547.

Bibliografía VII

[Lau+23] Hugo Laurençon et al. “The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset”. En: *CoRR* abs/2303.03915 (2023). DOI: 10.48550/arXiv.2303.03915. arXiv: 2303.03915. URL: https://doi.org/10.48550/arXiv.2303.03915.

[MHW23] Yasmin Moslem, Rejwanul Haque y Andy Way. “Adaptive Machine Translation with Large Language Models”. En: *ArXiv* abs/2301.13294 (2023).

Bibliografía VIII

[Ott+19] Myle Ott et al. “fairseq: A Fast, Extensible Toolkit for Sequence Modeling”. En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*. Ed. por Waleed Ammar, Annie Louis y Nasrin Mostafazadeh. Association for Computational Linguistics, 2019, págs. 48-53. DOI: 10.18653/v1/n19-4009. URL: https://doi.org/10.18653/v1/n19-4009.

Bibliografía IX

[Ram+20] Gema Ramírez-Sánchez et al. “Bifixer and Bicleaner: two open-source tools to clean your parallel data.”. En: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, nov. de 2020, págs. 291-298. ISBN: 978-989-33-0589-8.

[Rei+22] Ricardo Rei et al. “COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task”. En: *WMT*. Association for Computational Linguistics, 2022, págs. 578-585.

[Sca+22] Teven Le Scao et al. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”. En: *CoRR* abs/2211.05100 (2022). DOI: 10.48550/arXiv.2211.05100. arXiv: 2211.05100. URL: https://doi.org/10.48550/arXiv.2211.05100.

Bibliografía X

[Sha+17] Noam Shazeer et al. “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer”. En: *ICLR (Poster)*. OpenReview.net, 2017.

[Tie12] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. En: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. Ed. por Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2012, págs. 2214-2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/463.html.

Bibliografía XI

[Vas+17] Ashish Vaswani et al. “Attention is All you Need”. En: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. por Isabelle Guyon et al. 2017, págs. 5998-6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[Wan+21] Shuo Wang et al. “Language Models are Good Translators”. En: *CoRR* abs/2106.13627 (2021). arXiv: 2106.13627. URL: https://arxiv.org/abs/2106.13627.

Bibliografía XII

[Zha+23] Wayne Xin Zhao et al. “A Survey of Large Language Models”. En: *CoRR* abs/2303.18223 (2023). DOI: 10.48550/arXiv.2303.18223. arXiv: 2303.18223. URL: https://doi.org/10.48550/arXiv.2303.18223.

[ZHB23] Biao Zhang, Barry Haddow y Alexandra Birch. “Prompting Large Language Model for Machine Translation: A Case Study”. En: *ArXiv* abs/2301.07069 (2023).

[Zhu+23] Wenhao Zhu et al. “Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis”. En: *ArXiv* abs/2304.04675 (2023).