Multilingual Language Models

Linguistic Information in Large Language Models

Marion Di Marco

marion.dimarco@tum.de

July 2, 2024

Outline

- Large Language Models: state-of-the-art performance on many tasks
- Typically trained without explicit linguistic information,
  just large quantities of (multilingual) text
- Multilingual models: jointly trained on multiple languages,
  typically no explicit marking of the languages

⇒ Zero-shot cross lingual transfer in multilingual models

⇒ Multilingual capabilities in (English-centric) Large Language Models

⇒ Low-resource languages in LLMs

Outline

Multilingual Models and Cross-Lingual transfer

- Multilingual models have been shown to work surprisingly well for zero-shot cross-lingual transfer
  - Train a model on multiple languages
  - Fine-tune the model on a task in one language (typically English)
  - Apply the model to solve the task in another language
    multilingual pre-training → generalization to other languages
- Bridge the gap to lower-resourced languages

- mBERT: language model pre-trained from monolingual corpora in 104 languages
- Shared word piece vocabulary
- No direct cross-lingual supervision

Generalization Across Languages

- **How multilingual is Multilingual BERT?** Pires et al. (2019)
- Evidence that LMs such as BERT encode e.g. syntactic and named entity information
- To what degree generalize these representations across languages?

- Zero-shot cross-lingual model transfer with mBERT
  - supervised task-specific fine-tuning for language `A`
  - evaluate that task in language `B`
  - → analyze generalization of information across languages

Experiments and Results

- Named entity recognition

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

- Part-of-speech tagging

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

Tables from Pires et al. (2019)

Effect of Vocabulary Overlap

- Does transferability depend on lexical overlap ($\rightarrow$ vocabulary memorization)?
- Transfer to languages written in different scripts (no overlap)?

- Compute overlap of word pieces in the training and evaluation data
- Compare NER F1 scores for zero-shot transfer between every language pair of 16 languages for EN-BERT and M-BERT
  - EN-BERT: performance depends directly on word piece overlap
  - M-BERT: good performance even for lower overlap

  $\rightarrow$ representational capacity beyond simple vocabulary memorization

Effect of Vocabulary Overlap

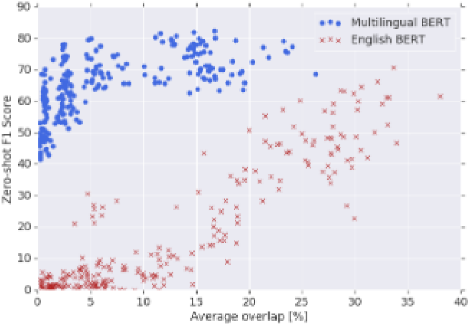


Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

Figure from Pires et al. (2019)

Generalization Across Scripts: POS tagging

- M-BERThas a surprising ability to transfer between languages written in different scripts (i.e. effectively zero lexical overlap)
- despite training on separate monolingual corpora without multilingual objective

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- High results between Urdu (Arabic script) and Hindi (Devanagari script)
- Less accurate for other pairs (e.g. EN – JA) → topological similarities

Table from Pires et al. (2019)

Effect of Language Similarity

- Comparison based on WALS features relevant to grammatical ordering

Figure 2: Zero-shot POS accuracy versus number of common WALS features. Due to their scarcity, we exclude pairs with no common features.

- Performance improves with language similarity → better mapping of linguistic structures for more similar languages

Figure from Pires et al. (2019)

Generalizing Across Typological Similarities

- POS accuracies for transfer between languages grouped according to two typological features:
  - subject/object/verb order
  - adjective/noun order
- Results reported include only zero-shot transfer

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

Table 5: Macro-average POS accuracies when transferring between SVO/SOV languages or AN/NA languages. Row = fine-tuning, column = evaluation.

- Best performance between languages sharing word order features,
  - ability to map learned structures onto new vocabularies,
  - less able to transfer structures to different word orders

Table from Pires et al. (2019)

Cross-Lingual Abilities of mBERT

- Hypothesis: Pires et al. (2019), Cao et al. (2020), Wu and Dredze (2019)

  cross-lingual abilities of mBERT are based on a combination of

  – (i) shared vocabulary items that act as anchor points;
  – (ii) joint training across multiple languages that spreads this effect; which ultimately yields
  – (iii) deep cross-lingual representations that generalize across languages and tasks

- Artetxe et al. (2020) take a closer look at this hypothesis :

  propose an alternative approach:

  cross-lingual transfer of monolingual representations

Cross-Lingual Transferability of Monolingual Representations

- **On the Cross-lingual Transferability of Monolingual Representations** Artetxe et al. (2020)

- Train a transformer-based masked LM on one language, then transfer it to a new language
- This approach does not rely on a shared vocabulary or joint training
- Competitive with multilingual BERT on standard cross-lingual classification benchmarks and on a new Cross-lingual Question Answering Dataset (XQuAD).

Cross-Lingual Transferability of Monolingual Representations

- L1: monolingual corpus and task supervision
- L2: only monolingual corpus
- Separate subword vocabulary for each language,

(1) Pre-train monolingual BERT in L1 (masked language modeling and next sentence prediction)

(2) Transfer model to a new language: learn new token embeddings on language L2 while freezing the transformer body

(3) Fine-tune the transformer for a task using labeled data in L1, while keeping the L1 token embeddings frozen

(4) Zero-shot transfer the resulting model to L2 by swapping the L1 token embeddings with the L2 embeddings

Models and Settings

- **Joint multilingual models** (JOINTMULTI): multilingual BERT model trained jointly on 15 languages
- **joint pairwise bilingual models** (JOINTPAIR): multilingual BERT model trained jointly on two languages (English and another language)
- **Cross-lingual transfer of monolingual models** (MONOTRANS): as described above; English as L1

- Vocabulary:
  - JOINTMULTI:
    models with a vocabulary of 32k, 64k, 100k, and 200k subwords
  - JOINTPAIR:
    model with a joint vocabulary of 32k (learned for each language pair);
    model with a disjoint vocabulary of 32k subwords per language (learned on the monolingual corpus, same vocab as MONOTRANS)

- 14 languages (fr, es, de, el, bg, ru, tr, ar, vi, th, zh, hi, sw, ur)

Experiments: XNLI (Natural Language Inference)

- NLI: given two sentences (a premise and a hypothesis), decide whether there is an entailment, contradiction, or neutral relationship

- JOINTMULTI is comparable with the literature
- Vocabulary: JOINTMULTI variants with larger vocabulary are better
- More languages do not improve performance. JOINTPAIR models with a joint vocabulary perform comparably with JOINTMULTI
- A shared subword vocabulary is not necessary for joint multilingual pre-training. JOINTPAIR models with a disjoint vocabulary for each language perform better
- MONOTRANS is competitive with joint learning. The best model variants are slightly worse than JOINTPAIR

Experiments – Summary

- Further experiments (document classification, paraphrase identification, question answering) → similar results
- **Joint multilingual training**
  - sharing subwords across languages is not necessary
  - no clear improvements by scaling to a large number of languages
  - effective vocabulary size per language is an important factor: joint vocabulary → only a subset is effectively shared
  - JOINTPAIR models with disjoint vocab generally perform best
- **Transfer of monolingual representations**
  - MONOTRANS is competitive even in challenging scenarios
  - suggests that multilingual pre-training is not essential for cross-lingual generalization
  - Probing the representations of MONOTRANS: monolingual models learn some semantic abstractions that are generalizable to other languages

Outline

Large Language Models – Languages

gpt-3 / dataset_statistics / languages_by_word_count.csv

language	number of words	percentage of total words
en	181014683608	92.64708%
fr	3553061536	1.81853%
de	2870869396	1.46937%
es	1510070974	0.77289%
it	1187784217	0.60793%
pt	1025413869	0.52483%
nl	669055061	0.34244%
ru	368157074	0.18843%
ro	308182352	0.15773%
pl	303812362	0.15550%
fi	221644679	0.11344%
da	221551540	0.11339%
sv	220920577	0.11307%
ja	217047918	0.11109%

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

Large Language Models – Languages

- Many Large Language models are English-centric

Table 1 Top five languages included in GPT-3 training data compared against other measures of the top five global languages, from 1st most common and widely used.

	1st	2nd	3rd	4th	5th
GPT-3 training data (2019) [35]	English (93%)	French (1.8%),	German (1.5%)	Spanish (0.8%)	Italian (0.6%)
Languages represented on the Internet (2021) [36]	English (44.9%)	Russian (7.2%)	German (5.9%)	Chinese languages (4.6%)	Japanese (4.5%)
First-languages spoken (2019) [37]	Mandarin Chinese (12%)	Spanish (6%),	English (5%),	Hindi (4.4%),	Bengali (4%).
Most spoken language (2021)[37]	English (1348M)	Mandarin Chinese (1120M)	Hindi (600M)	Spanish (543M)	Standard Arabic (274M)

Figure from Johnson et al. (2022)

Multilingual Capabilities of Large-Scale LMs

- **On the Multilingual Capabilities of Very Large-Scale English Language Models** Armengol-Estapé et al. (2022)

- LLMs are predominantly English → multilingual capabilities?

- Large majority (93 %) of GPT-3's training data is English
- Comparatively small portions of other languages
- Is this enough for good LMs in those languages?

Multilingual Capabilities of Large-Scale LMs

- Previous work: focus mostly on capabilities for tasks in English
- MT with GPT-3: good for translating *into* English

- Evaluate GPT-3 on 3 generative tasks
  - extractive question-answering,
  - text summarization,
  - natural language generation
  - 5 languages: German, Spanish, Russian, Turkish, Catalan
  - different model sizes

Zero-Shot Multilingual Question Answering

- Question Answering: produce an answer given a context and a question
- XQuAD: benchmark dataset for evaluating crosslingual QA performance — Artetxe et al., (2020)
  - subset of SQuAD translated into ten languages — Rajpurkar et al., (2016)
  - same question+answer pairs for all languages
    → no bias wrt. difficulty
- Example

  This is a Question-Answering system in English.

  Context: The Panthers defense gave up just 308 points [...]

  Question: How many points did the Panthers defense surrender?

  Answer: 308
- Prompts are formulated in the evaluated language

Zero-Shot Multilingual Question Answering

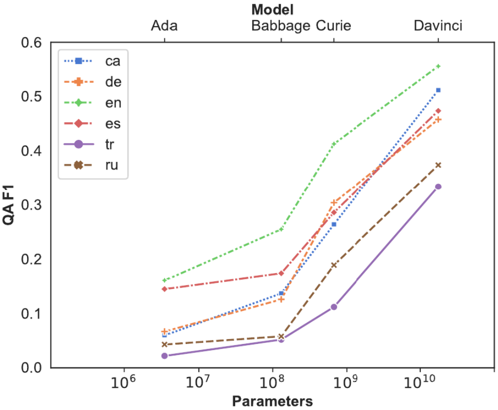


Figure 1: Automatic metrics results (F1) for the question-answering task

Figure from Armengol-Estapé et al. (2022)

Zero-Shot Multilingual Text Summarization

- Producing a shorter version of a text while preserving relevant information
- MLSUM: a multilingual summarization dataset obtained from online newspapers — Scialom et al. (2020)
  - multilingual content is **not** parallel
  - Catalan: CaSum dataset (manually revised)
- Filtering
  - length: text + summary + instruction exceeds context window (2048 tokens)
  - quality: summaries with a ROUGE score below 0.1
  - Russian: discarded entirely (→ English-centric tokenization)
- Prompt format:

  [... text ...] TL;DR

Zero-Shot Multilingual Text Summarization: Evaluation

- Generation tasks are difficult to evaluate
- Words in the summary ↔ words in the reference
- Length: how long is a good summary?
  - in supervised learning: similar length as in examples
  - (→ zero-shot setting in the experiment)
  - in the used data set: most summaries are not longer than 3 sentences

- ROUGE: N-gram co-occurrences
- Manual evaluation for EN and CA (ranking)

Zero-Shot Multilingual Text Summarization: Evaluation

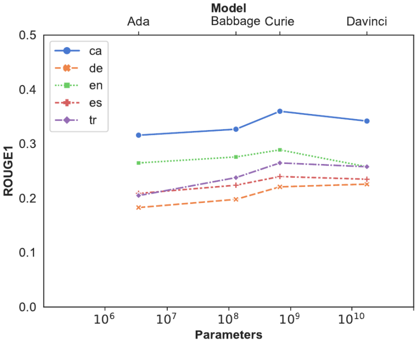


Figure 2: Automatic metrics results (ROUGE-1) for the Text Summarization task

- Davinci: random manual inspection

  more concise summaries, more creative in terms of the lexical choices

Figure from Armengol-Estapé et al. (2022)

Zero-Shot Multilingual Text Summarization: Evaluation

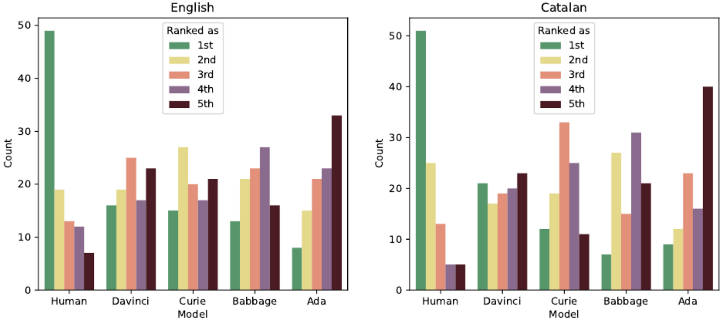


Figure 3: Human ranking results for the Text Summarization Evaluation task.

Figure from Armengol-Estapé et al. (2022)

Zero-Shot Multilingual Text Generation

- Turing test: was a sentence produced by a human or by AI?
- High cost of human evaluation: only Catalan and English

- Data set: randomly sample 20 news articles and use the headline as prompt
- Generate text in the same language as the headline
- Select 60 sentences each from the generated articles and the original articles
- 3 native speakers decide: human or AI generated $\Rightarrow$ majority vote

Zero-Shot Multilingual Text Summarization: Evaluation

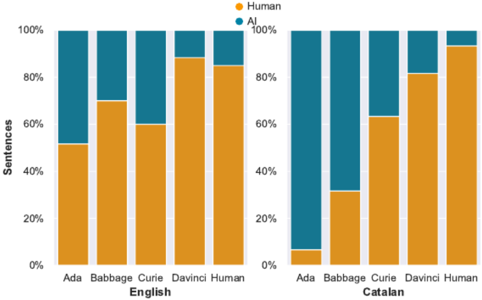


Figure 4: Human evaluation results for the Text Generation task

- Inter-annotator agreement:
  Fleiss $\kappa = 0.401$ for Catalan and 0.290 for English

Figure from Armengol-Estapé et al. (2022)

Multilingual Capabilities of Large-Scale LMs: Discussion

- Remarkable zero-shot generative capabilities in languages that appear in tiny proportions in the training corpus
  - Russian: non-Latin alphabet
  - Turkish: no typological affiliation
  - Catalan: moderately under-resourced

- Scaling: transfer learning between English and the other languages in zero-shot settings scales with model size

- Tokenization: English-based segmentation
  - token/word ratio as a predictor for GPT-3 performance
  - Russian: excluded from summary task due to segmentation

⇒ GPT-3: almost as useful for many languages as it is for English

Outline

Monolingual or Multilingual LLMs?

- **Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models** Blevins et al. (2022)

- Many LLMs are presented as *English* models, but have been found to transfer well to other languages
- Common English pre-training corpora contain significant amounts of non-English text

  Even a small percentage → hundreds of millions of foreign language tokens in large-scale datasets
- Small percentages of non-English data facilitate cross-lingual transfer with the performance strongly correlated to the amount of in-language data

Non-English Data

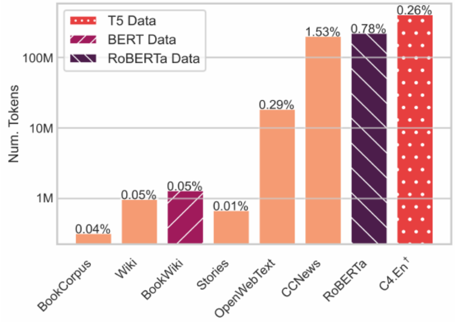


Figure 1: Estimated non-English data in English pre-training corpora (token count and total percentage); even small percentages lead to many tokens. C4.En (†) is estimated from the first 50M examples in the corpus.

Figure from Blevins et al. (2022)

Non-English Data

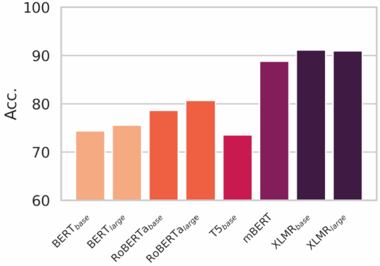
Type	Book	Wiki	Stories	OpenWeb	CCNews	C4
	Num. of Lines in...					
NE	156	129	99	175	193	169
	Ex: Moraliska argument utgår ifrån våra moraliska intuitioner att rätt och fel inte endast är förankrade i människors vilja. (OPENWEBTEXT)					
BiL	13	11	15	4	1	22
	Ex: The German blazon reads: "Von Silber über Schwarz geteilt..." (WIKI)					
Trans.	2	7	4	2	0	4
	Ex: Εκείνη δεν μπορούσε να πληρώσει [She couldn't pay.] (BOOKCORPUS)					
Ent.	1	28	5	1	0	1
	Ex: 2012 Playhouse Presents ウィルシリーズ1、エピソード1: "The Minor Character" (C4)					
En	26	22	55	12	6	3
	Ex: "Dere's buzzards circlin' ova dem trees." (BOOKCORPUS)					
XX	2	3	22	6	0	1
	Ex: M D I X O X I O O O = A (WIKI)					

Table 1: Results of the qualitative analysis of the non-English lines in various pretraining corpora. Type abbreviations are defined in Section 2.2.

Figure from Blevins et al. (2022)

Experiment: POS Probing

- Train linear classifier to predict POS from the final layer of the encoder

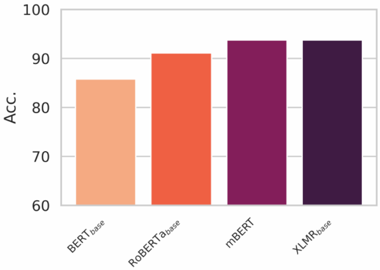


(b) POS (probing)

  - T5: more absolute non-English data than RoBERTa, but less in terms of relative percentage (0.78% vs. 0.22%)
  - RoBERTa's subword tokenization is more robust than T5 and BERT
  - For many high-resource languages: English models perform competitively; T5 outperforms mBERT for German and Portuguese

Experiment: POS Fine-tuning

- Fine-tuning for non-English POS-tagging



(c) POS (finetuned)

  - Gap between the mono- and multilingual models becomes smaller
  - RoBERTa averages 2.65 points worse than XLM-R, compared to 12.5 points when probing

Outline

Under-Represented Languages

- There are ≈ 7000 languages in the world
- A majority of languages is not represented in pre-trained LMs

- mBERT, multilingual roBERTa: ≈ 100 languages
- GPT-3: 119 languages listed (last position: *Cham* with 49 words)
- NLLB (No Language Left Behind): translation model for 200 languages

  Costa-Jussà et al. (2022)

- Glot500: 511 languages — Imani et al. (2023)
  - skewed distribution of languages
  - "head" languages: comparatively large languages
  - "tail" languages: smaller languages with little to no resources

Under-Represented Languages: Data

- The performance of a language model is dependent on training data in the target language
- Adapt the pretrained multilingual models to low-resource languages?
- Constrained by the amount of monolingual or parallel data available
  → difficult for languages with little or no textual data

- Language documentation: bilingual lexicons or word lists

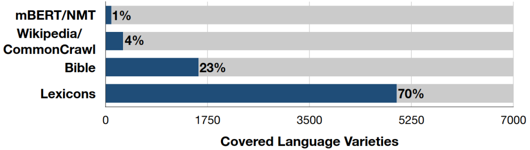


Figure 1: The percentage of the world's ≈7,000 languages covered by mBERT, monolingual data sources and lexicons.

Figure from Wang et al. (2022)

Learning Endangered Languages with Linguistic Descriptions

- **Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions** Zhang et al. (2024)
- Idea: put targeted linguistic knowledge into the prompt

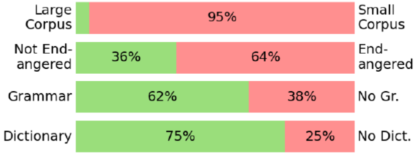


Figure 1: Among the world's ∼7000 languages, 95% don't have enough data (>100K sentences) for training LLMs (Bapna et al., 2022), while most have a grammar book (60%) or dictionary (75%) (Nordhoff and Hammarström, 2011), including many endangered languages (Moseley, 2010). Therefore, we utilize these linguistic descriptions to bring LLMs to endangered languages.

Learning Endangered Languages with Linguistic Descriptions

- How does a linguist analyze an utterance in a foreign language?
  $\Rightarrow$ Dictionary and grammar book!

- Most languages have some linguistic resources
- Linguistic descriptions are different from text collections:
  - Smaller in size
  - Instructional: explicit grammar rules that can be used as instructions for both LLMs and humans
- Dictionary and or grammar book: too large for the prompt context
  $\Rightarrow$ exploit available linguistic resources to handle languages unseen in pre-training

Incorporating Linguistic Descriptions

(1) Morphological Analysis: Source Sentence $\rightarrow$ Morphemes
   - (existing) finite-state morphological analyzers

(2) Dictionary Mapping: Morphemes $\rightarrow$ Gloss
   - language dependent: words vs. stems
   - lookup in a dictionary,
     strategies to handle no/multiple matches (e.g. edit distance)

(3) Incorporating Grammar Knowledge: Gloss $\rightarrow$ Translation and Beyond
   - Some word-level grammatical information is already covered in the morpological analysis
   - Prompt the LM with grammar knowledge
     (some pre-processing required)

Incorporating Linguistic Descriptions

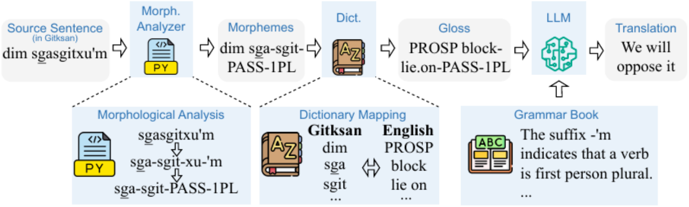


Figure 3: LINGOLLM uses a morphological analyzer to transform the source sentence into morphemes, looks up the morphemes in a dictionary to obtain the gloss, and finally feeds both the gloss and a grammar book to an LLM to obtain the result.

Figure from Zhang et al. (2022)

Incorporating Linguistic Descriptions: Experiments

- 8 typologically and geographically diverse endangered or low-resource languages
- 5 tasks: translation from/to English, mathematical reasoning, response selection, word reordering, and keyword-to-text

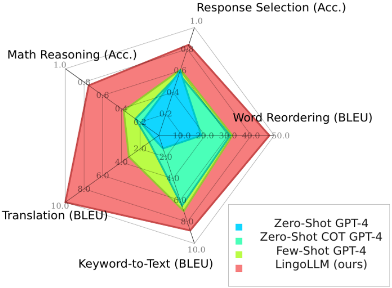


Figure 2: LINGOLLM significantly outperforms GPT-4 on 5 NLP tasks across 8 endangered or low-resource languages.

Figure from Zhang et al. (2022)

Outline

mBERT: Cross-Lingual Transfer

Multilingual Capabilities of Large-Scale LMs

Monolingual or Multilingual LLMs?

Low-Resource and Endangered Languages in LLMs

Summary

References

Summary

- Cross-Lingual Transfer in mBERT: relevant features
- Large-scale LMs: multilingual capabilities with
- Languages represented in LLMs: English vs. Non-English
- Strategies to model low-resourced languages in LLMs

Outline

References

- Telmo Pires, Eva Schlinger, Dan Garrette (2019):
  How multilingual is Multilingual BERT?
  Proceedings of ACL 2019. `https://aclanthology.org/P19-1493.pdf`
- Steven Cao, Nikita Kitaev, Dan Klein (2020):
  Multilingual Alignment of Contextual Word Representations.
  ICLR 2020. `https://arxiv.org/abs/2002.03518`
- Shijie Wu, Mark Dredze (2019):
  Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT.
  EMNPL 2019. `https://aclanthology.org/D19-1077/`
- Mikel Artetxe, Sebastian Ruder, Dani Yogatama (2020):
  On the Cross-lingual Transferability of Monolingual Representations
  Proceedings of ACL 2020. `https://aclanthology.org/2020.acl-main.421.pdf`
- Rebecca L. Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, Donald Jay Bertulfo (2022):
  The Ghost in the Machine has an American accent: value conflict in GPT-3
  `https://arxiv.org/abs/2203.07785`

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019):
  BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  Proceedings of NAACL 2019. `https://aclanthology.org/N19-1423/`
- Yinhan Liu, Myle Ott et al. (2019):
  RoBERTa: A Robustly Optimized BERT Pretraining Approach
  `https://arxiv.org/abs/1907.11692`
- Jordi Armengol-Estapé, Ona de Gibert Bonet, Maite Melero (2022):
  On the Multilingual Capabilities of Very Large-Scale English Language Models.
  LREC 2022. `https://aclanthology.org/2022.lrec-1.327/`
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang (2016):
  SQuAD: 100,000+ Questions for Machine Comprehension of Text
  Proceedings of EMNLP 2016. `https://aclanthology.org/D16-1264/`
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano (2020): *MLSUM: The Multilingual Summarization Corpus*.
  EMNLP 2020. `https://aclanthology.org/2020.emnlp-main.647/`
- Alexis Conneau, Guillaume Lample et al. (2018):
  XNLI: Evaluating Cross-lingual Sentence Representations
  `https://arxiv.org/abs/1809.05053`

References

- Terra Blevins, Luke Zettlemoyer (2022).
  Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models. EMNLP 2022.
  `https://aclanthology.org/2022.emnlp-main.233/`
- NLLB-Team, Marta R. Costa-Jussà et a. (2022):
  No Language Left Behind: Scaling Human-Centered Machine Translation.
  `https://arxiv.org/abs/2207.04672`
- Ayyoob ImaniGooghari et al. (2023):
  Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages.
  ACL 2023. `https://aclanthology.org/2023.acl-long.61/`
- Xinyi Wang, Sebastian Ruder, Graham Neubig (2022):
  Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation.
  ACL 2022. `https://aclanthology.org/2022.acl-long.61.pdf`
- Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, Lei Li (2024): *Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions.* `https://arxiv.org/abs/2402.18025`