Cross-lingual (Visual) Language Models on Understanding Physical Concepts

Shengqiang Zhang, Yihong Liu

October 7, 2024

1. Outline: Cross-Lingual NLP
2. Multilingual Language Models
3. Cross-lingual Transfer Learning

Cross-Lingual NLP involves developing systems that can **understand** and **generate** text in **multiple languages**.

This capability is crucial in a globalized world where information needs to be accessible **across linguistic boundaries**.

- **Communication**: Facilitates communication in multilingual settings.
- **Information Access**: Provides access to information in low-resource languages.
- **Cultural Exchange**: Promotes understanding and exchange between cultures.

Challenges in Cross-Lingual NLP:

- **Language Diversity**: Over 7,000 languages worldwide, with significant grammatical and syntactic differences.
- **Resource Scarcity**: Most languages lack large annotated datasets, which are essential for training machine learning models.
- **Ambiguity**: Words or phrases may have different meanings in different languages (e.g., "bank" in English; "chat" in English and French).
- **Syntax and Grammar Differences**: Sentence structures vary greatly, making it challenging to maintain coherence across languages.

Typical Applications:

- **Machine Translation**: Converting text from one language to another, e.g., Google Translate.
- **Cross-Lingual Information Retrieval**: Retrieving information in one language based on a query in another.
- **Multilingual Sentiment Analysis**: Analyzing sentiment in social media posts across various languages.
- **Cross-Language Dialogue Systems**: Building chatbots that can interact with users in multiple languages.

Key Concepts in Cross-Lingual NLP

Languages and Scripts:

- **Language Families**: Languages grouped by common ancestry, such as Indo-European, Sino-Tibetan, and Afro-Asiatic.
- **Scripts**: Writing systems used by different languages, like Latin (English, Spanish), Cyrillic (Russian, Bulgarian), and Arabic (Arabic, Persian).

Key Concepts in Cross-Lingual NLP

Linguistic Divergences:

- **Phonetic Differences**: Variations in sounds and pronunciation.
  - Example: “th" sound in English is not present in many languages.
- **Syntactic Differences**: Variations in sentence structure and grammar.
  - Example: Subject-Verb-Object (SVO) in English vs. Subject-Object-Verb (SOV) in Japanese.
- **Semantic Differences**: Words may have different meanings or nuances.
  - Example: The word “gift" means “present" in English but “poison" in German.

Data Resources:

- **Parallel Corpora**: Bilingual text pairs used for training machine translation models (e.g., Europarl corpus).
- **Comparable Corpora**: Non-parallel, but related texts in different languages (e.g., Wikipedia articles on the same topic).
- **Bilingual Dictionaries**: Lists of word translations between languages, useful for basic cross-lingual tasks.

Outline

Definition, Scope, and Advantages:

- **Multilingual Models**: Language models that are designed to process and understand text in multiple languages using a single architecture.
- **Scope**: They can handle tasks like translation, sentiment analysis, and question answering across different languages (usually requiring fine-tuning on downstream tasks).
- **Advantages**:
  - **Efficiency** – single model for multiple languages
  - **Transfer** – high-resource languages benefit the low-resource ones
  - **Flexibility** – adaptable to new languages

Multilingual Embeddings:

- **Word Embeddings**:
  - **Approach-I**: first learn monolingual embeddings for multiple languages and then align them in a shared vector space (Artetxe et al., 2017; Lample et al., 2018; Artetxe et al., 2018).
  - **Approach-II**: directly learn multilingual word embeddings for all languages by creating special data structures, e.g., graph (Dufter et al., 2018; Liu et al., 2023).
- **Sentence / Document Emebeddings**:

  Similar to word embeddings, sentence-level or document-level embeddings can be obtained by taking the average of the word embeddings or additionally combining techniques such as TF-IDF.

Transformer-based Models

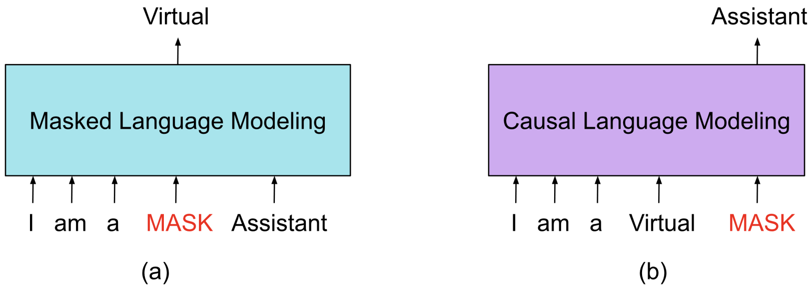
- **Encoder-only Models**: mBERT, XLM-R, XLM-V, IndicBERT, AfriBERTa, Glot500-m ...
  - Typically good at language understanding tasks, e.g., sentiment analysis.
- **Decoder-only Models**: XGLM, mGPT, BLOOM, Llama, PaLM, GPT (3, 3.5, 4), PolyLM, Aya ...
  - Typically good at language generation tasks, e.g., story generation.
- **Encoder-Decoder Models**: mBART and mT5 ...
  - Typically good at controlled generation tasks, e.g., machine translation.

Data Preparation

- **Gathering Multilingual Datasets**: Use resources like Common Crawl, Wikipedia, and genre-specific corpora like the Bible.
- **Data Cleaning**: Remove noise or redundancy and ensure consistency in multilingual datasets.
- **Data Augmentation** (optional but beneficial): Techniques like back-translation and synthetic data generation to enhance low-resource language data.

Training Objectives

- **Masked Language Modeling**
- **Causal Language Modeling**
- ...



Downstream Tasks & Benchmarks

- **XTREME** (Cross-lingual TRansfer Evaluation of Multilingual Encoders): Covers tasks like text classification, QA, and translation across languages (Hu et al., 2020).
- **XGLUE** (Cross-lingual General Language Understanding Evaluation): Evaluates models on tasks like NER, QA, and text classification across languages (Liang et al., 2020).
- ...

Rule of thumb: Evaluate in a way that can evaluate the crosslinguality of the multilingual model.

Outline

Cross-lingual transfer learning aims to leverage knowledge from high-resource languages to improve NLP tasks in low-resource languages.

Advantages:

- **Resource Efficiency**: Reduces the need for large datasets in every language.
- **Accelerated Model Development**: Speeds up the development process for new languages.
- **Performance Enhancement**: Improves accuracy and robustness for low-resource language tasks.

Types of Transfer Learning

- **Zero-Shot Learning**: Model performs tasks in a new language without explicit training data for that language.
  - **Example**: Training on English and directly applying to Spanish without Spanish data.
- **Few-Shot Learning**: Model adapts to a new language with minimal training examples.
- **Transfer Across Related Languages**: Leveraging similarities in related languages (e.g., Spanish and Portuguese) for better transfer learning.

Techniques for Cross-Lingual Transfer Learning

- **Continued Pretraining**: Use models trained on multilingual data as a starting point and continually pretrain it on new languages (Wang et al., 2022; Alabi et al., 2022; ImaniGooghari et al., 2023).
  - **Example**: Continually pretrain XLM-R on 500 languages -> Glot500-m.
- **Knowledge Distillation**: Large, complex models (teachers) transfer knowledge to smaller, simpler models (students) for specific languages (Jiao et al., 2020; Sanh et al., 2020).
  - **Example**: Distilling knowledge from mBERT to a smaller model optimized for a particular language.
- **Cross-lingual Alignment**: Aligning word or sentence embeddings across languages using parallel corpora, bilingual dictionaries, or even monolingual corpora (Gao et al., 2021; Zhang et al., 2023; Liu et al., 2024).
  - **Example**: Contrastive learning to improve the similarity between matched pairs (words/sentences) against random pairs.

Thank you for your attention!

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 generalization. *CoRR*, abs/2003.11080.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schuetze. 2023. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8376–8401, Singapore. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Junlei Zhang, Zhenzhong Lan, and Junxian He. 2023. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3916–3932, Singapore. Association for Computational Linguistics.