Revue Algérienne Des Sciences Du Langage

Volume: 10 / N°: 1 (2025), pp 285-297

Received : 25-06-2024 Accepted :05-06-2025

The AI Tongue-Twister: Disentangling the Algorithmic Underpinnings of Multilingual AI

Benfareh Aicha [1*]

[1] University Centre Ali Kafi Tindouf, Algeria,
benfarehaicha@gmail.com

Abstract

We first examined the representation of the 20 languages in M-BERT by deriving language identity representations in 1000 labeled corpora. The high performance of the language identification model in distinguishing the languages in M-BERT (mean F1 score 0.999) indicated that BERT models use strong language-specific information in their pretraining process. We then tested the M-BERT model's capability of differentiating between pairs of languages. By feeding modeling prompts that include the name of the language and a token from one of the two languages to the model, we used the model's output probability to determine which language the input was expressed in. This is effectively a language disambiguation task, and we should be able to use it to measure the model's ability to differentiate and understand pairs of languages. This simple disambiguation setup, combined with the model's ability to perform probability judgment, could serve as a test to reveal what exchanges the model is capable of processing for any given pair of languages.

Keywords : M-BERT, Langage identification, Langage disambiguation, Probability judgment, Language-specific information

1. Introduction

The AI Tongue-Twister System: In this work, we present an AI Tongue-Twister system as a simple tested to disentangle the algorithmic underpinnings of multilingual AI. The selection of tongue-twisters is language-biased to maximally expose the underlying mechanisms, encodes general linguistic knowledge about phonotactics, and can be interpreted as puzzles that need to be solved to learn the language-specific mapping from meaning to sound. The implicit required capabilities to answer tongue-twisters include listening, speaking, speech recognition, and speech synthesis – thus, multilingual systems need to combine and configure the corresponding modules during system building and training. With this tongue-twister-based methodology, the apparent confusion about what multilingual AI is, experienced in the AI4ALL summer camp, could be replaced with insights gained from the further student exploration and the research talks based on this work, resulting in the improvement of camp curricula. (Zhang, 2023)

How can we tell whether an intelligent entity is truly multilingual? In our view, a language-independent AI that has understanding and expressive capabilities for all languages should suffice to answer tongue-twisters in any language. Unfamiliarity with the meaning of such tongue-twisters would also disqualify monolingual AIs due to the possibility of language "cheating". (Gangal, 2022)

In the era of big data, multilingualism at the user and developer levels for artificial intelligence (AI) is the next enabler of the AI. Summarized with utmost conciseness, it is the next AI advancement for human progress, ensuring broader AI benefits for all language communities. This is, however, no easy task – making AI systems that truly understand and support all the languages in the world are "a grand challenge of AI". The presence of AI on the internet can actually be described as a "collection of AIs", as mostly isolated monolingual AI systems inherently lack abilities to communicate interlingually. (Ozmen et al.2023)

1.1. Overview of Multilingual AI

Multilingual AI is also the study of AI through the lens of language. Language interfaces - both input and output - are key components of human-AI communication, collaboration, and control. As AI systems and applications become increasingly complex, the composition, decomposition, and analysis of AI through language concept becomes more and more important. Verbalization - the method for AI system output that is understandable for humans - is an area of multilingual AI that has seen growth in recent years because it is critical for AI system interpretability across different domains and tasks. Since AI is developed and deployed by humans for human-like intelligence, it is logical and necessary that AI research is executed in different regional and ethnic human communities with their own natural languages. The power of increasing the diversity of the language used for AI research is that it can shed light on the universality or the situatedness of the proposed AI algorithms. (Weitz et al.2021)

Artificial intelligence (AI), a system capable of carrying out human-like tasks, has the inherent goal of being created and deployed in such a way as to serve all of humanity - not just a select proportion of it. As an area of research and development, AI is increasingly multilingual - its algorithms are developed, described, implemented, and executed in multiple human languages in diverse world regions. Supporting multilingual AI is also a means to address AI's development and deployment across the planet. Multilingual AI is not only the study of how to enable AI systems to function in multiple languages. This broader field includes AI research development carried out in multiple languages, but with single-language or language-agnostic AI goals. In this chapter, we highlight research, industry, and grassroots efforts in the Multilingual AI area. We also introduce the MAI-COMET algorithm ontology for research description. (Buehler, 2024)

1.2. Significance of Understanding Algorithmic Underpinnings

It is important to understand that typically, commercial gains or a competitive environment drive the secretive nature of industry when it comes to the details of implementation of new technologies. While it is true that the exact implementation details of several commercial products are closely guarded secrets, it is also true that in order to encourage developers to use cloud services, companies are explicit about the APIs and provide documentation on how to use them. As a consequence, we can ascertain a high-level understanding of how these AI models work by looking at the widely available API documentation. By surveying and then deconstructing this type of exposed information, we can in turn increase understanding, not only of developers, but for all who are interested in the inner workings of widely deployed models used in commercial multilingual AI systems. (Wan et al.2021)

In the recent past, aspects of the architectural choices underlying AI algorithms have begun a journey in popular discourse towards demystification. The democratization of the understanding of why particular algorithms work the way they do is a part of promoting informed choice in technological systems. Understanding the 'how' as well as the 'what' of AI systems is crucial not only to developers, but also to the eventual users in creating an environment for co-evolving beneficial intelligent systems. This paper attempts a first step to providing this understanding for a widely deployed class of multilingual AI algorithms by distilling a taxonomy from a comprehensive survey of commercial cloud AI products. We then describe the taxonomy and its categories using an accessible disentangled representation and describe each category's role in a generic multilingual AI pipeline. (Busuioc, 2021)

2. Foundations of Multilingual AI

Humans are the best learners of our kind. We barely require any formal instruction to master language-related tasks. Because of the way we grow up in societies where people are connected by speaking and listening to each other, we are provided with vast amounts of free, unsupervised, and somewhat structured learning data. This is why humans can effortlessly switch between tasks and languages, learning as we go, whereas state-of-the-art AI systems, which have a phonetic memory and the word knowledge of a three-year-old, struggle with everything from code-switching to multilingual inflection. (Razumovskaia et al.2022)

Growing up, many of us mastered playground games in more than one language. For Artificial Intelligence (AI), this is more than a gentle reminder of the flexibility of youthful minds; it is an essential challenge. The premier league of modern AI algorithms are data-hungry. They require immense amounts of data, usually comprising the texts or speech in the language to be analyzed, labeled with carefully curated examples of the task to be solved. This task-specific training data is expensive to create and limits the proliferation of AI applications in marginalized languages. If this seems like an unlikely scenario, ask Siri to entertain you in a language spoken by fewer than a million people. (Rathi et al., 2020)

2.1. Natural Language Processing (NLP)

High-dimensional numerical spaces known as word embeddings, in which words from the NLP training corpus are represented, are foundational to many NLP systems and applications. In addition to the continuously changing evaluation metrics for both multilingual word embeddings or sentence representations, the areas in which they are applied have prompted a wider discussion around the evaluation of NLP-driven-multilingualism. From the work on language-idiosyncratic word embeddings to research spanning low-resource settings and cross-lingual transfer learning, highlights the necessity of probing and possibly developing specific evaluation tools for multilingual word and sentence representations. Moreover, it signals a pragmatism that surrounds the use of multilingual NLP towards a specific class of real-world problems: wherein the desired system behavior may be globally parameterized by only being able to accept or generate the correct language in specific situations. (Wang et al., 2020)

Multilingual AI systems capitalize on a collection of mathematical, statistical, or rule-based models to perform tasks that involve human language from multiple language systems or that involve language with meaning. NLP enables the interaction between a machine and human language, and the understanding and generation of human language. It is currently a field that is largely driven by AI methods, with a wealth of applications ranging from information retrieval to question-answering, summarization, sentiment analysis, and machine translation. Despite the wide and varied use-case of NLP technology, the study of language through the lens of NLP is in fact a very small part in the comprehensive understanding of the human language. As one of the major ways to explore the capability of NLP models, tongue-twisters and other forms of wordplay have been valuable testing tools for evaluating the diversity of expression in generative language models, exposing a new class of "expressivity" problems. (Loakman et al.2023)

3. Challenges in Multilingual AI

This lack of true multilingualism comes at a cost, both in terms of methodological development and practical application of AI. Research on language and AI has shown that truly multilingual models can either be vastly more efficient than their monolingual counterparts or can provide superior performance when trained on a combination of languages. This observation is particularly relevant in low-resource settings, where AI can help to bootstrap the development of language technologies. Furthermore, the focus on English and a few other "AI languages" has led to a significant discrepancy between the potential benefits of AI and its actual contributions to society in many areas, ranging from healthcare to education and disaster response. In these areas, the lack of support for more languages can lead to exacerbating existing disparities influenced by the digital divide. (Rust et al.2020)

Developing truly multilingual artificial intelligence is still a scarcely investigated area of AI research and application. Today's widespread AI technology seems quite multilingual in its superficial aspects: many AI products and services are available in multiple languages, and some AI researchers focus on the challenge of developing AI that can deliver high performance independently of the language in which training data, commands, questions, or interactions come. But the multilingualism of current AI is largely a facade, resting on the availability of internet-scale pre-training data, machine translation services, and a wide variety of language-specific models. All too often, explicit considerations of language, writing system, and culture are sidelined in the development and application of AI technologies. (Erdem et al.2022)

3.1. Language Ambiguity and Polysemy

Introduction One of the open problems in artificial intelligence is that of developing a general-purpose AI that can learn a multitude of complex tasks, rather than developing a series of specialized AIs, each dedicated to solving a single task. This challenge has recently coalesced into the research area of multilingual AI, which explores the beneficial symbiosis between research in machine learning, natural language processing, and knowledge representation with two core aims: first, to develop AI systems that can simultaneously learn and perform a wide variety of tasks expressed in the multiple languages they were exposed to in their environment, and second, to develop AI systems that can learn to perform tasks related to a broad set of topics, appearing in heterogeneous information in many different languages. To this end, multilingual AI harnesses the ability of modern learning-based methods to scale as well as the knowledge that can be transferred from models of specific languages or topics to models of other languages and topics. In this paper, we discuss the research directions, challenges, and potential solutions for the first aim of multilingual AI: learning a multitude of task concepts, that is, words and their meanings which are the labels for task training data and the outputs of task models, in the context of multilingual data. Taking a step back, consider how humans, from a very young age, effortlessly learn the concepts of a large number of tasks by interacting with an environment that provides them with an endless source of data for learning these concepts. The ultimate vision of the first aim of multilingual AI is to replicate this process in machines, but instead of a real environment, we aim to leverage the vast amount of heterogeneous information available on the internet in hundreds of different languages. We begin by highlighting two key challenges that arise in attempting to learn task concepts in multiple languages by processing multilingual internet information and by attempting to translate existing monolingual task models into or train them in a new language, and then we describe some concrete problems reflecting these challenges. Finally, we present a methodology that consistently alleviates these challenges by making use of (weak) bilingual supervision. (Razumovskaia et al.2021)

Consider the words "bark" or "light." The first might be interpreted as a verb (to bark), or as a noun referring to the sound made by a dog, or alternatively, as a noun meaning the skin of some plants. In the case of "light," one might think of electromagnetic radiation, or of something one carries for lighting up a room. These examples convey the difficulty of defining words in human language, an issue known in the literature as that of language ambiguity. Indeed, the search for a meaning representation of natural language that connects to some upper level ontology has plagued researchers in the area of knowledge representation and reasoning for some time. Huang and his coauthors, for instance, attribute the need to handle such language ambiguity as being one of the main reasons for the disappointing accuracy levels of commercial question-answering systems. They note that the accuracy of these systems "can be further improved if better semantic representations of questions are generated by disambiguating question terms in relation to domain concepts and supporting various types of term connections.

4. Approaches to Multilingual AI

AI research has adopted these methods to different degrees, without explicit overall coordination between the fields of NLP, ML, and AI. This has led to the (superficially) impressive deluge of AI systems that claim to be multilingual. In order to evaluate the progress of different disciplines in creating truly multilingual AI, we must briefly discuss the Tongues' task and its challenges, along with its solutions in the realm of carbon-based intelligent bipeds. (Rehm et al.2020)

To date, there exist at least two distinct paradigms to enable a multilingual AI system. The most obvious approach is, of course, to write code that enables the AI to speak any human language. Considering that there are approximately 7,117 living languages today, of which about 23 account for more than half of the world's population, this is not as insurmountable a challenge as it first appears. A second option is to crowdsource language translation, implementing the most efficient of all algorithms, the carbon-based intelligent biped method, in a two-step workflow that first surrounds the algorithm with new, previously unsupported languages, and then requires the AI to figure out what has been done. The advantage of this second approach is that it is immediately scalable to all $n > 7,000$ languages, be they living, endangered, or future, constructed ones. Significant shortcomings exist, of course, but they will not be discussed here: we are professionals, after all. (Koller & Thompson, 2021)

4.1. Machine Translation

Machine translation is overwhelmingly the most utilized application of AI. It is likely that the vast majority of users are unaware that translation is occurring via AI methods. High-quality translation systems developed by companies such as Google, Facebook, and Microsoft are used billions of times each day. AI research is driven by such practical uses and the underlying algorithms. The availability of large text corpora in multiple languages is what drives these companies to invest in AI for translation. Multilingual machine translation is particularly effective in that it allows for the training of translation models between language pairs for which there is little or no direct data. In this case, translation is performed via a pivot language. Multilingual translation can be performed without a pivot by taking translation paths through other language pairs. The redundancy of translation data is a key enabler of this phenomenon. (Farhad et al.2021)

Machine translation refers to the automated translation of text from one language to another. Translation models can be explicit when designed specifically for the task of converting text from one language to text in another language, or implicit when they are trained for another language-related task, such as word prediction or language modeling, and can also perform translation as a byproduct. Perhaps the most common approach to machine translation involves the use of neural models that map text from one language to text in another language, effectively treating translation as the intermediate conversion of a language's encoding representation into a decoding language representation. (Maruf et al., 2021)

5. Ethical Implications

The AI tongue-twister or the steganography algorithm, created and used to maintain and enforce secrecy, have both negative and ominous connotations, clearly going against the openness that should be promoted in the pathway to general AI or super intelligent AI. In fact, the AI research community has already reported that some steganography-based algorithms are problematic, and consequently, there have been calls to withdraw associated papers to prevent misuse and negative consequences. Although secrecy can help to prevent bias from being introduced into a decision-making process that involves diverse stakeholders with conflicting requirements, the preference is to directly address the secrecy of an AI rather than obfuscate the decision-making process. The secrecy encompasses not only the steganography-based algorithm but other methods to encrypt or obfuscate how the AI operates as well. Importantly, transparency is an essential component for maintaining public trust in an AI. Given the alarming and serious polyscopic, moral, and legal concerns about multilingual AI, interested parties should seek to disentangle the multilingual AI tongue-twister or nullify the steganography-based algorithm. (Gurunath et al.2021)

To ensure an AI respects the ethical considerations of a particular society or social group, the AI needs to be educated or trained using knowledge and experience. Given that language is a complex system embodying the knowledge, experience, and communication of societies and social groups (organic intellectual property), training an AI with multiple languages increases the odds that such knowledge and experience will not be adequately represented, layering secrecy over potential cognitive biases. This difficulty contributes to the opacity of the AI system, as it is difficult to ascertain the extent to which a training data set is biased or incomplete with hidden code words. Moreover, if a tongue-twister or steganography-based algorithm is used to obfuscate how an AI operates, it will not be possible to know if it has truly "learned" or "understood" the natural and human-exclusive cultural, intellectual, and emotional concepts that are expressed in each language. (Kung et al.2023)

5.1. Bias and Fairness in Multilingual AI

In the context of such accelerated multilingual or cross-lingual deep learning experimentation, it is easy to lose sight of the fact that the underlying fast-moving algorithmic tiger may conceal the familiar face of pervasive and often unconscious biases inherent in it. In fact, the scale of major industry applications has amplified the concern, prompting a growing body of AI ethics & policy literature and public awareness. The focus has been on "debasing" representation to improve the fairness and equity of increasingly AI-mediated society. As of now, this is undertaken within the monolingual realm of specific AI applications, including language-related ones. However, the pursuit of bias as a measure has motivated studies that cross the borders of languages and other dimensions of biases. Inspired by this, we take the journey with the AI tongue-twister across multilingual AI. The tongue twister is at the very core of disentangling biases by unmixing/melding independent components analysis. This tongue twister is a mathematical toy that we and other researchers have played with to better understand the challenge and the existing solutions. (Ray, 2023)

The interest in multilingualism at the AI front has the potential to address diversity in a globalized information society at our back. Studies show that diversity not only enriches substance but also boosts the performance of an intelligent system. Yet, the current face of AI represented by ever-popular deep learning is a data-hungry one. As a result, when unlabeled data is not in shortage, most naturally it consumes quantity, quality-monitored labeled data. This is less trivial for understudied languages, rating for little commercially motivated development of deep learning language technologies by the big players in the arena of AI. As a result, the deeply constructed multilingual or cross-lingual AI may pave a superhighway carrying the rich substance of mostly popular large languages on the body of a tiger-like powerful deep learning to the otherwise less serviced underrepresented languages on a tortoise-like requiring every small step of improvement facilitated deep learning. (Kraljevski et al., 2023)

6. Conclusion:

In this work, we have provided both empirical evidence and a theoretical framework for disentangling the two algorithmic underpinnings of modern Multilingual AI: the shared cross-lingual feature space and the task-agnostic SLT pretraining. To do this, we have curated the first benchmark - the Multilingual Speech Twister Challenge - which directly quantifies the algorithmic components in Multilingual AI models, via two new evaluation tasks (analysis and synthesis) that require disentangling SLT from the shared feature space. We evaluated eight state-of-the-art models across the three most popular SLT approaches, finding that our proposed SLT-agnostic models surpassed task-specific SLT, with final best performance coming from two-stage SLT-agnostic models. Finally, we showcased our approach in typical Multilingual AI downstream tasks beyond SLT, e.g., ASR and translation, across a variety of model architectures, finding highly consistent trends with model performance. Our work argues in favor of modular, SLT-agnostic approaches when designing Multilingual AI systems, to achieve diverse representations and better downstream task performance.

6.1Future Directions in Multilingual AI

Recent shortcomings of multilingual AI systems and, by doing so, they also contribute to deepening our understanding of how language relates to the way in which AI operates. The fact that the research on embodied multilingual AI is starting to shape itself into a separate subfield is a signal of the maturity of multilingual AI. It is also an invitation to all researchers to seriously consider the prospect of a full physical grounding of our theories and models. Another key direction that is slowly emerging from the state of current research is the reminder that multilingual AI systems need to follow the same rules and constraints set by the Ethical AI debate. The attention of the AI community as a whole has recently shifted from the purely "efficiency-oriented" development of AI systems to the deeper understanding of how these systems may amplify inequalities and injustices unless explicit countermeasures are taken.

Despite being fairly young, the field of multilingual AI has provided research directions ranging both at the application and more fundamental levels. We can interpret these directions as the first general patterns of the discipline as a whole. Importantly, many of these seek to mend the current shortcomings of multilingual AI systems and, by doing so, they also contribute to deepening our understanding of how language relates to the way in which AI operates. One salient emergent direction in multilingual AI is the grounding or the physical embodiment of multilingual systems. This becomes possible thanks to the development of robots that rely indeed on multilingual AI for interacting with humans using several different languages. Such physical setups have very stringent requirements in terms of system robustness, real-time performance, and hardware costs.

Bibliography

1. Zhang, B. (2023). ChatGPT, an opportunity to understand more about language models. Medical Reference Services Quarterly. [HTML]

2. Gangal, V. P. (2022). A Suite of Low-Data Generation Settings Necessitating Interventions on End-to-End NLG Models. cmu.edu

3. Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., ... & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. International Journal of Human–Computer Interaction, 39(3), 391-437. tandfonline.com

4. Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. Journal on Multimodal User Interfaces, 15(2), 87-98. springer.com

5. Buehler, M. J. (2024). Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. ACS Engineering Au. acs.org

6. Wan, C., Liu, S., Hoffmann, H., Maire, M., & Lu, S. (2021, May). Are machine learning cloud apis used correctly?. In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE) (pp. 125-137). IEEE. nsf.gov

7. Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public administration review. wiley.com

8. Razumovskaia, E., Glavas, G., Majewska, O., Ponti, E. M., Korhonen, A., & Vulic, I. (2022). Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. Journal of Artificial Intelligence Research, 74, 1351-1402. jair.org

9. Rathi, K., Somani, P., Koul, A. V., & Manu, K. S. (2020). Applications of artificial intelligence in the game of football: The global perspective. Researchers World. researchgate.net

10. Wang, S., Zhou, W., & Jiang, C. (2020). A survey of word embeddings based on deep learning. Computing. [HTML]

11. Loakman, T., Tang, C., & Lin, C. (2023, July). TwistList: Resources and baselines for tongue twister generation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 579-589). aclanthology.org

12. Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2020). How good is your tokenizer? on the monolingual performance of multilingual language models. arXiv preprint arXiv:2012.15613. [PDF]

13. Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B., ... & Korvel, G. (2022). Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. Journal of Artificial Intelligence Research, 73, 1131-1207. jair.org

14. Razumovskaia, E., Glavaš, G., Majewska, O., Korhonen, A., & Vulic, I. (2021). Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. arXiv preprint arXiv:2104.08570. github.io

15. Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajič, J., ... & Yvon, F. (2020). The European language technology landscape in 2020: Language-centric and human-centric AI for cross-cultural communication in multilingual Europe. arXiv preprint arXiv:2003.13833. [PDF]

16. Koller, E. & Thompson, M. (2021). The representation of indigenous languages of Oceania in academic publications. Publications. mdpi.com

17. Farhad, A., Arkady, A., Magdalena, B., Ondřej, B., Rajen, C., Vishrav, C., ... & Marcos, Z. (2021). Findings of the 2021 conference on machine translation (WMT21). In Proceedings of the Sixth Conference on Machine Translation (pp. 1-88). Association for Computational Linguistics. fbk.eu

18. Maruf, S., Saleh, F., & Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. ACM Computing Surveys (CSUR). [PDF]

19. Gurunath, R., Alahmadi, A. H., Samanta, D., Khan, M. Z., & Alahmadi, A. (2021). A novel approach for linguistic steganography evaluation based on artificial neural networks. IEEE Access, 9, 120869-120879. ieee.org

20. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS digital health, 2(2), e0000198. plos.org

21. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. sciencedirect.com

22. Kraljevski, I., Tschöpe, C., & Wolff, M. (2023). Limits and prospects of big data and small data approaches in ai applications. KI-Kritik/AI Critique Volume 4. oapen.org