Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241249

Artificial intelligence's role in the realm of endangered languages: Documentation and teaching

Luyi Wang

Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA 90089, United States.

louiswan@usc.edu

Abstract. With numerous languages nearing extinction, the urgency to preserve endangered languages has become a prominent focus in the linguistic field. This paper delves into the transformative role of Artificial Intelligence (AI) in the domains of documentation and pedagogy for endangered languages, particularly highlighting its innovative applications and the associated challenges. It delves into how AI-powered tools reshape linguistic fieldwork, offering accelerated annotation, consistent data collection, and deeper analytical endeavors. Furthermore, this exploration highlights the potential of AI in revolutionizing the teaching of these languages, ushering in a new era marked by dynamic, scalable, and engaging learning experiences. While AI presents unparalleled efficiencies, its challenges, ranging from data scarcity to the looming digital divide, are addressed critically. As the digital age continues to evolve, merging AI's capabilities with traditional linguistic approaches holds the promise of a more inclusive and comprehensive strategy to rejuvenate and preserve the world's rich linguistic tapestry. This paper has summarized and provided an outlook on the research topic at hand.

Keywords: Endangered Languages, Artificial Intelligence (AI), Linguistic Fieldwork, Documentation, Pedagogy.

1. Introduction

The tapestry of languages worldwide is under threat, with many unique linguistic threads at risk of fraying into obscurity. Austin & Sallabank underscore the gravity of this situation, revealing the intricacies of beliefs and ideologies intertwined with language documentation and revitalization [1]. These endangered languages don't just encapsulate a unique way of communicating but are repositories of rich cultural traditions and worldviews.

Yet, with challenges come opportunities. The dawn of the digital age has ushered in an era of innovation where technology intersects with linguistics, aiming to stem the tide of linguistic erosion. Notably, the role of Artificial Intelligence (AI) is emerging as a beacon of hope. AI techniques are instrumental in the linguistic decolonization process, aiding the revival of endangered tongues. This shift isn't just about harnessing new tools; it's about reimagining the methodologies of linguistic fieldwork in the modern era. Zhang, Frey, & Bansal showcased that deep learning—a powerful AI subset—has the potential to push the boundaries and be pivotal in both documenting and rejuvenating endangered languages [2].

As endangered languages teeter on the brink of extinction, there's more at stake than mere words. They embody centuries of heritage, stories, and wisdom. However, current documentation methodologies may not be robust enough to counter this decline effectively. This raises a pertinent question: How can we optimize the blend of tradition and technology to safeguard linguistic diversity? This research aims to address this gap, spotlighting AI as a prospective solution. This article delves into the tangible impacts of AI in documenting and teaching endangered languages. The initial focus is on the pivotal role of AI technologies in capturing and preserving endangered languages, highlighting its invaluable contributions in the construction of grammar and vocabulary databases. Subsequently, the discourse transitions to how AI tools and platforms can be adeptly utilized for teaching endangered languages, showcasing actual application cases. However, while the benefits of AI in this context are undeniable, the article also critically examines some of the limitations associated with its use. This study delves into how AI aids in recording, preserving, and imparting endangered languages while also addressing its potential drawbacks.

2. Artificial Intelligence and the Documentation of Endangered Languages

In the quest to document endangered languages, Artificial Intelligence (AI) emerges as an indispensable ally. Its profound intersection with linguistic fieldwork reshapes traditional methodologies, offering innovative solutions.

2.1. Annotation and AI: Reshaping Linguistic Documentations

Annotation tools signify the beginning of this synergy. Annotation technology has traditionally been rooted in manual methods. Textual annotations involved simple techniques such as highlighting or tagging, while image annotations saw users drawing on or around areas of interest. Videos brought in time-bound tags, capturing significant moments, and audio annotations in linguistics enabled detailed tagging of specific sounds or pitches. A notable advancement was interactive and collaborative annotation, where multiple users could add collective insights to shared documents.

Linguistic annotations, whether they be morphological, syntactic, or semantic, undergo a transformative shift when processed by AI-driven tools. Textual annotations leveraged Natural Language Processing to automate entity and sentiment identification. Computer vision algorithms transformed image and video annotations, auto-detecting objects, tracking motion, and discerning facial expressions. Audio annotations were enhanced to distinguish multiple voices and predict muffled content. AI-backed tools even began offering predictive annotations, using past data to suggest potential areas of interest. For instance, recognized for its high programmability, serves as a pivotal tool for annotators. While it has significantly accelerated the annotation process, particularly vital for endangered languages, its integration with AI remains less seamless, limiting its full potential for auto-suggestions and other AI-driven features [3,4].

Moving a step beyond traditional tools, the Maori Language App serves as a beacon for endangered language revitalization. Designed specifically for the Maori language of New Zealand, this app incorporates AI to recognize speech patterns and provide real-time feedback, while its annotations unravel sentence structures, morphemes, and phonetic nuances, acting as an invaluable aid to learners.

2.2. Data Collection: AI as the Backbone

While tools like the Maori Language App have revolutionized annotations, advancements in AI extend even further, particularly in the realm of data collection for languages that traditionally lacked extensive resources. Traditional methods, often manual, relied heavily on surveys, manual transcription, and direct observations. With the advent of AI, however, these processes have been supercharged. Natural Language Processing (NLP) algorithms, for instance, can now scan vast amounts of text, extracting and categorizing relevant information in fractions of the time once required.

This paradigm shift is evident in tools like Lig-Aikuma, a mobile application highlighted by Blachon et al., specifically tailored for concurrent speech collection in the study of under-resourced languages

[5]. AI here is not a mere facilitator; it’s a guarantor of data quality, consistency, and uniformity, pillars for comprehensive linguistic analysis.

Transitioning from individual applications to broader initiatives, the Universal Dependencies (UD) project, emerges as a frontrunner. While the primary canvas of UD isn’t restricted to lesser-known languages, its expansive and inclusive approach has been gradually incorporating data from these linguistic subsets. Empowered by AI, UD excels in addressing linguistic intricacies and granularities, providing a consistent annotation approach across a myriad of languages (Nivre et al., 2019) [4].

Building on this theme of championing under-represented languages, the tech world has seen further innovation in more consumer-focused applications. A sterling example of AI’s dominant role in data collection is the recent development of voice-enabled assistants that cater to less-common languages. These assistants, such as specific versions of Siri or Alexa, utilize vast and deep learning algorithms to understand, interpret, and respond in numerous languages, some of which were previously under-documented. While these platforms assist users in their day-to-day tasks and serve as massive data collection hubs, they are not without concerns. As every interaction potentially refines the AI’s understanding of the language, capturing nuances, idioms, and colloquialisms, there arise pressing questions about user privacy. Often, these assistants are always listening, and inadvertent recordings can sometimes be stored and analyzed. The balance between linguistic advancement and safeguarding user privacy remains a point of contention.

2.3. AI in Analytical Linguistic Endeavors

AI further augments analytical endeavors. Machine learning and deep learning models have expanded linguistic analyses beyond traditional confines, unlocking unprecedented depth. As Neubig et al. [6] highlighted, AI algorithms can spot intricate linguistic patterns even in nascent documentation phases.

An exemplar of such prowess, the DeepSBD project utilized deep learning for a longstanding linguistic dilemma: segmenting and marking sentence boundaries in transcriptions, which underscored AI’s proficiency in demarcating sentences amidst continuous speech, adapting to the linguistic intricacies inherent across languages.

Moreover, it’s the groundbreaking transformer models like BERT and GPT that stand as testament to AI’s analytical potential. Upon introduction, BERT achieved state-of-the-art results on 11 NLP tasks, outstripping previous bests, such as scoring 80.5% on the GLUE benchmark—a measure of models on diverse tasks from question answering to sentiment analysis [7-8]. Meanwhile, GPT-2 presented unparalleled text generation, achieving a Cross-Entropy score of 18.34 on the Penn Treebank dataset [9].

Adding a theoretical layer, Goldsmith and O’Brien spotlighted unsupervised learning models’ significance in linguistic evaluation. Such models, capable of discerning patterns and structures sans pre-existing labels, become invaluable for analyzing lesser-known languages [10].

Quantitative successes—like BERT’s benchmark achievements and GPT’s text generation results—alongside the evolving theoretical stances affirm AI’s transformative impact in linguistic analysis, fostering a richer, more nuanced comprehension of diverse languages.

3. Challenges in Documentation

3.1. Sparse Data for Endangered Languages

While AI’s foray into linguistics offers great promise, it stumbles upon a major impediment: the limited data available for endangered languages. Languages like English and Chinese benefit from a plethora of digital datasets, facilitating robust AI model training. Conversely, endangered languages grapple with a digital resource scarcity, which complicates tailored AI model development. It’s noteworthy that these lesser-studied languages could greatly benefit from AI, especially when human expertise is limited. AI systems could operate as substitutes, offering insights and assistance. Yet, this potential boon is hampered by the inherent data limitations these languages face.

3.2. Lack of Pre-existing Models for Adaptation

Major languages benefit from pre-existing models ready for adaptation, whereas linguists working with less-studied languages often must start from scratch. The computational cost of training language models from scratch can be prohibitive. Strubell et al. highlighted that training one iteration of a large model like BERT might consume as much electricity as an average American household does in over two months [11]. This not only translates to heightened financial burdens but also elongates the research and development timeline, posing further challenges in the race to document and revitalize endangered languages.

3.3. Phonetic, Orthographic, and Dialectal Complexities

Endangered languages, with their myriad phonetic, orthographic, and dialectal variations, often throw a curveball at standard AI tools conditioned on mainstream linguistic benchmarks. These unique characteristics, while integral to the cultural identity of these languages, sometimes befuddle AI systems accustomed to more prevalent linguistic standards.

3.4. Ethical and Cultural Concerns

Languages are more than communication tools; they're cultural repositories. As AI ventures into documenting languages, significant ethical dilemmas surface. The primary concern hinges on cultural sensitivity: Can AI, like Siri or Alexa, truly capture a language's essence without unintentionally distorting its cultural nuances? Moreover, the act of data collection, especially from indigenous groups, raises questions about consent and transparency. While AI's role in linguistics promises innovation, it simultaneously demands a vigilant ethical approach.

However, solutions emerge. Specifically, for languages with extremely limited labeled data - much like endangered languages - methods such as few-shot and zero-shot learning offer a beacon of hope. These strategies empower AI models to deliver informed predictions even when presented with a meager number of examples. To further tackle the issue of data scarcity, data augmentation strategies come into play. By artificially enlarging the training dataset using methods like back-translation or synonym substitution, there's an opportunity to diversify training examples and thus enhance model efficacy for endangered languages.

4. Artificial Intelligence in the Teaching of Endangered Languages

4.1. Laying the Groundwork for AI-Infused Pedagogy

Transitioning from documentation, the pedagogical domain witnesses a transformative wave with AI, particularly in the context of endangered languages. Insights from Zhang, Frey, & Bansal on the Cherokee language are testament to AI's burgeoning role in education [2]. With AI at the helm, plenty of innovative tools, including digital platforms, interactive games, and advanced translation mechanisms, have emerged. These not only captivate the digital-native generation but also make learning these rare languages more accessible. This chapter aims to explore the diverse avenues through which AI is revolutionizing the pedagogical approach to endangered languages.

4.2. NLP-Powered Quiz Generation: Bridging the Resource Gap

Endangered languages often face a lack of adequate teaching resources. Here, Natural Language Processing (NLP) offers a promising solution. It enables the autonomous crafting of quizzes, from cloze tests to multiple-choice questions. Moreover, the inclusion of gamification elements, like crosswords and flashcards, enhances student engagement and aids retention. Recent studies underscore the potential of NLP in educational settings. For instance, Heilman and Smith discussed an NLP system that automatically generates cloze test items, a development that can significantly benefit endangered language education by streamlining resource creation [12]. Furthermore, Burstein et al. highlighted the use of NLP tools in creating multiple-choice questions from textual content [13]. Gamification, as a pedagogical strategy, has garnered attention too. Hamari, Koivisto, & Sarsa emphasized the effectiveness of gamification elements, like crosswords and flashcards, in enhancing student motivation and learning outcomes [14].

4.3. The Dawn of Automated Assessment: A New Age of Feedback

Historically, language assessment has largely depended on manual evaluations by educators. While this approach offers a personal touch, it faces drawbacks. Grading inconsistencies due to biases, the lengthy assessment process, and delayed feedback can impede learning. Additionally, ensuring uniformity and scalability for diverse learners and languages poses challenges.

In stark contrast, the advent of the digital era has propelled the field of automated language assessment forward. Evaluating students' grasp of endangered languages becomes streamlined with automated assessments. While intricate evaluations, such as grading essays, might still pose challenges, AI can efficiently handle foundational assessments. Tools that provide real-time feedback, analyzing linguistic parameters like word frequency and syntactical intricacies, bolster the learning trajectory. However, it's crucial to note that foundational NLP instruments like Part-Of-Speech (POS) taggers, remain indispensable for achieving this vision. In this context, Shermis and Hamner highlighted the capabilities of automated essay scoring systems, even though they underscored the complexities involved in fine-tuning them for nuanced evaluations [15]. Tetreault and Chodorow have illustrated the role of NLP tools, specifically POS taggers, in diagnosing writing errors, underlining their importance in real-time feedback systems [16]. Leveraging such tools for endangered languages might require adaptations, but the foundational methodologies remain relevant.

4.4. Embracing Community-based Learning: Collaboration is Key

Traditional classroom settings and tutor-style instruction for endangered languages are often restrictive in scalability, unable to cater to a wider audience. Additionally, they may lack immediacy in feedback and adaptability to individual learning paces, making it challenging for learners to consistently engage and progress.

Shifting to contemporary solutions, Duolingo stands out as a model for the effective incorporation of AI in language pedagogy. The platform's AI-driven lessons personalize learning experiences based on individual progress and performance, thereby addressing the adaptability concern inherent in traditional settings. Furthermore, Duolingo's spaced repetition algorithms ensure that learners review concepts at optimal intervals, enhancing long-term retention. Its gamified approach, underpinned by AI, not only makes learning engaging but also provides instantaneous feedback, which is crucial for astering a new language.

While Duolingo and platforms like Tatoeba, which focuses on sentence-based translation, have set the stage, there's potential to further leverage AI for endangered languages. Envision platforms tailored for these languages, bringing together native speakers, linguists, and learners. Such integrative digital ecosystems could significantly contribute to the revitalization and preservation of endangered languages.

5. Navigating the Pitfalls in Technological Pedagogy

5.1. Over-reliance on AI and Loss of Human Interaction

While AI tools in language education enhance learning, there's a risk of learners becoming excessively reliant on them. Such an over-reliance could overshadow the essential human touchpoints in language learning, particularly in grasping conversational subtleties and cultural nuances [17].

5.2. Marginalization of Dialectal Variations

AI tools tend to gravitate towards standardized forms of languages, which may inadvertently overshadow or even neglect vital dialectal variations. This skewness threatens the rich tapestry of linguistic diversity, pushing the essence of local dialects and variants to the periphery.

5.3. Technological Divide and Equity Concerns

The introduction of AI in language pedagogy brings forth the specter of the digital divide. Not every learner has access to state-of-the-art technological tools, leading to an uneven playing field. This divide further intensifies pre-existing educational disparities, placing those without access at a significant disadvantage [17].

5.4. Lack of Emotional Support and Motivation

While AI tools can efficiently tackle the technical facets of a language, they fall short in providing the emotional and motivational support intrinsic to human educators. The encouragement, understanding, and connection a human teacher can offer are irreplaceable, especially during challenging phases of language learning [18].

To counter these challenges, a blended learning approach, combining AI tools with traditional teaching methods, is essential. By integrating technology with human interaction, learners can benefit from AI's efficiency while preserving the invaluable nuances of face-to-face communication.

6. Conclusion

This exploration delved deeply into the symbiotic relationship between Artificial Intelligence and endangered languages. Confronted with the intricacies of documenting these languages, AI, despite grappling with challenges such as sparse data, absence of existing models, and linguistic complexities, sees the dawn of hope with innovative approaches like zero-shot and transfer learning. In the educational sphere, AI's transformative capacity shines through. While traditional classroom and tutor-led pedagogies hold value, they are constricted in terms of scalability and adaptability. In contrast, modern platforms like Duolingo harness AI to pioneer an era marked by dynamic and scalable learning experiences. However, a note of caution resonates: while AI-augmented tools offer unparalleled efficiency, they cannot supplant the irreplaceable nuances only human touch can provide.

Looking to the future, the merger of AI with endangered language studies exudes immense potential. The evolution of large language models (LLMs) is pivotal in this regard. Marrying AI's efficiencies with the profound depths of traditional methodologies, we stand on the precipice of a more inclusive, vibrant, and comprehensive strategy to breathe life into and preserve the rich tapestry of the world's linguistic heritage.

References

[1] Austin, P. K., & Sallabank, J. (2014). Endangered Languages: An Introduction. Oxford University Press.

[2] Zhang, S., Frey, B., &amp; Bansal, M. (2022). How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[3] Wittenburg, P., Brugman, H., & Russel, A. et al. (2006). "ELAN: A Professional Framework for Multimodality Research." Proceedings of LREC.

[4] Nivre, J., de Marneffe, M. C., & Ginter, F. et al. (2019). "Universal Dependencies 2.4." LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[5] Blachon, D., Hamlaoui, F., & Jacobson, O. et al. (2016). "Parallel Speech Collection for Under-Resourced Language Studies Using the Lig-Aikuma Mobile Device App." Procedia Computer Science, 81, 61-66.

[6] Neubig, G., Duh, K., & Sudoh, K. (2011). "Learning to Translate with Multiple Objectives." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

[7] Wang, A., Singh, A., & Michael, J. et al. (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." Proceedings of the 2018 EMNLP Conference.

[8] Devlin, J., Chang, M. W., & Lee, K. et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.

[9] Radford, A., Wu, J., & Child, R. et al. (2019). “Language Models are Unsupervised Multitask Learners.” OpenAI Blog.

[10] Goldsmith, J., & O’Brien, D. (2006). “Learning Inflectional Classes.” Language, 82(3), 555-592.

[11] Strubell, E., Ganesh, A., &amp; McCallum, A. (2019). Energy and policy considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[12] Heilman, M., & Smith, N. A. (2010). “Good Question! Statistical Ranking for Question Generation.” Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[13] Burstein, J., Chodorow, M., & Leacock, C. (2015). “Automatic Essay Evaluation: The Criterion Online Writing Service.” AI Magazine, 27(3).

[14] Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? -- A literature review of empirical studies on Gamification. 2014 47th Hawaii International Conference on System Sciences.

[15] Shermis, M. D., & Hamner, B. (2013). Automated Essay Scoring: A Cross-disciplinary Perspective. Routledge.

[16] Tetreault, J. R., & Chodorow, M. (2008). “The Ups and Downs of Preposition Error Detection in ESL Writing.” Proceedings of Coling.

[17] Warschauer, M. (2003). Technology and Social Inclusion: Rethinking the Digital Divide. MIT Press.

[18] Dörnyei, Z., & Ushioda, E. (2011). Teaching and Researching Motivation (2nd ed.). Pearson Education.