Large Language Models and Low-Resource Languages: An Examination of Armenian NLP

Hayastan Avetisyan and **David Broneske**
The German Centre for Higher Education Research and Science Studies (DZHW), Germany
avetisyan@dzhw.eu
broneske@dzhw.eu

Abstract

This paper presents a comprehensive review of Natural Language Processing (NLP) research on Armenian, a language that, despite its rich history and unique linguistic characteristics, is currently low-resource in the field of NLP. We critically synthesize and evaluate various studies in Armenian NLP, highlighting key advancements, challenges, and areas for improvement. A notable aspect of our work is the underlined lack of application of Large Language Models (LLMs) in Armenian NLP, signifying an area of potential exploration and development. Identifying and discussing these challenges and opportunities lays the groundwork for future research directions in Armenian NLP. The emphasis on Armenian also advocates for increased attention to low-resource languages in NLP research, stressing the importance of linguistic diversity and equity. To the best of our knowledge, this is the first paper providing such an extensive review of Armenian NLP, marking a significant contribution to the field.

1 Introduction

The evolution of Natural Language Processing (NLP) technologies has revolutionized our interactions with digital systems, paving the way for more human-like dialogues with artificial intelligence. Despite these advancements, a significant number of the world's languages are still considered low-resource in terms of NLP tools and technologies. One such language is Armenian, an Indo-European language with unique complexities and limited digital resources. To address these issues and shed light on the Armenian language in the context of NLP, we present a comprehensive study that contributes to the field in several ways:

1. **Comprehensive Overview of Armenian NLP**: We provide a thorough review of the current state of Armenian Natural Language Processing (NLP), identifying key research, significant advancements, and areas of improvement within the field. We synthesize and critically evaluate the findings from various studies, providing valuable insights for researchers and practitioners alike.

2. **Identifying Challenges and Opportunities**: Our paper highlights the challenges of developing NLP technologies for Armenian, a low-resource language. It elucidates the issues related to the lack of linguistic resources and tools, gaps in machine learning methodologies, and the specific challenges associated with Armenian. This knowledge is essential for future research and resource allocation.

3. **Emphasis on Large Language Models (LLMs)**: Our paper underscores the underutilization of LLMs in Armenian NLP. It examines the potential benefits and challenges of deploying these models in low-resource settings and suggests avenues for future research.

4. **Future Directions**: We present an informed and thorough projection of potential future work in Armenian NLP, covering various aspects such as the development of linguistic resources, the application of machine learning techniques, and the need for multilingual and cross-lingual models. This helps set a clear roadmap for other researchers in the field.

5. **Encouragement of NLP Research for Low-Resource Languages and Promotion of Language Diversity**: By focusing on Armenian, a low-resource language, our paper advocates for more attention to be given to similar languages in NLP research. It stresses the importance of these languages in ensuring linguistic diversity and equity in NLP.

To the best of our knowledge, this is the first comprehensive review focusing specifically on the

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 199–210

advancements, challenges, and opportunities in Armenian Natural Language Processing. Therefore, our paper fills a crucial gap in the literature, providing a much-needed foundation for further exploration and development in Armenian NLP. This added context further emphasizes the value and novelty of our contributions to the field.

2 Exploration of Armenian Natural Language Processing: Key Domains and Literature Review

In the forthcoming chapter, we will present an overview of the current state of Armenian NLP based on the findings of our review. Given the scope limitations, a detailed discussion of the findings cannot be provided. However, for a comprehensive understanding of the literature search process, including the specific databases searched and the keywords employed, and to ensure transparency, please refer to Section A in the appendix.

2.1 Linguistic Resources and Tools

The rapid growth of Natural Language Processing (NLP) has primarily been geared toward resource-rich languages such as English, Spanish, and Chinese, overlooking lower-resource languages like Armenian. This oversight results in a scarcity of linguistic resources and tools, which is detrimental to language preservation, linguistic research, and the development of NLP applications for such languages (Silberztein, 2007; Baghdasaryan, 2022; Malajyan et al., 2020; Kindt and Van Elverdinghe, 2022; Megerdoomian, 2009; Vidal-Gorène and Decours-Perez, 2020; Khurshudyan et al., 2022; Saini and Rakholia, 2016).

NLP Development Platforms: NooJ, for example, allows users to create grammars and dictionaries while providing parsing capabilities for tokenization, morphology, syntax, and semantics. However, its resources are still insufficient for full-scale Armenian language processing, indicating a need for more comprehensive platforms (Silberztein, 2007).

Speech Corpora: The ArmSpeech corpus presents a collection of Armenian speech data, which is essential for speech recognition, synthesis, and spoken language identification systems. Nevertheless, the necessity for similar large-scale resources in other aspects of the Armenian language remains (Baghdasaryan, 2022).

Semantic and Paraphrase Corpora: The ARPA corpus, a paraphrase dataset generated using back translation, forms a foundation for paraphrase detection in Armenian. However, the small scale of this corpus (2360 paraphrases) underscores the need for more diverse datasets to enhance NLP tasks (Malajyan et al., 2020).

Historical Language Resources: Efforts like the GREgORI Project, which processes historical Armenian manuscripts, are crucial for enhancing Armenian linguistic resources. However, a broader range of Armenian language variations and contexts needs to be included for comprehensive resource development (Kindt and Van Elverdinghe, 2022).

Language-specific Features and Bootstrapping: Techniques such as related language bootstrapping can help develop language resources. However, Armenian's unique linguistic features necessitate a language-specific approach rather than a generalized strategy (Megerdoomian, 2009).

Resource Compilation Platforms: Calfa compiles and updates existing resources for Classical Armenian, showcasing promising advancements in Armenian NLP. However, the need for larger, interoperable databases persists, indicating the necessity for more systematic efforts (Vidal-Gorène and Decours-Perez, 2020).

Comprehensive Corpora: The Eastern Armenian National Corpus (EANC) provides a valuable resource for linguistic studies and various research fields. However, the need for more comprehensive databases covering different time frames, styles, and registers is apparent (Khurshudyan et al., 2022).

Foundational Resources: Specific stop-word lists for Armenian are notably lacking, despite being foundational resources in NLP (Saini and Rakholia, 2016).

In conclusion, while progress has been made, further efforts are required to enrich Armenian NLP. This necessitates the creation of more comprehensive resources, the development of tools for diverse NLP tasks, and the promotion of multilingual or cross-lingual learning methods to counterbalance data scarcity. Moreover, it highlights the need for collaboration among researchers, data scientists, and the Armenian community worldwide to create culturally sensitive, community-driven resources.

2.2 Evolution of Speech Processing and Recognition

Armenian speech processing and recognition has witnessed transformative advancements across various key sub-domains, primarily automated speech recognition (ASR), speech recognition thresholds (SRT), dialect identification, and evaluation of ASR systems (Davit and Tigran, 2022; Baghdasaryan; Sargsyan and Rahne, 2021; Avetisyan, 2022; Aslanyan et al., 2015; Chakmakjian and Wang, 2022).

Automated Speech Recognition (ASR) Models: A dual strategy has marked Progress in this realm. Davit and Tigran (2022) fine-tuned a pre-trained Conformer model yielding notable Armenian ASR outcomes, while Baghdasaryan devised an efficient real-time speech-to-text system using Baidu's Deep Speech and KenLM toolkit. These initiatives signal the feasible extension of ASR applications to languages with limited resources.

Speech Recognition Thresholds: Sargsyan and Rahne (2021) innovated a multisyllabic speech audiometry test capable of measuring SRT in Armenian speakers. The results corroborated with equivalent studies in other languages, signifying a universal scope for this approach.

Dialect Identification: The work of Avetisyan (2022) illustrated the superiority of neural network models over statistical approaches in Armenian dialect identification. The successful utilization of pre-trained word vectors, even with scarce training data, added to the novelty of their work.

Evaluation and Future Directions for ASR: Aslanyan et al. (2015) meticulously assessed current ASR technologies and suggested alternatives to Hidden Markov Models, like Artificial Neural Networks and Sequence Data Mining, to enhance recognition accuracy.

Challenges of ASR for Bivariant Languages: The complexities of developing ASR for Armenian, a language with two diverse standard variants, were highlighted by Chakmakjian and Wang (2022). A practical methodology emphasizing the need for a comprehensive understanding of both language variants was proposed for future ASR research.

These studies underscore the advances in Armenian speech processing and recognition, mapping out promising future directions for low-resource languages like Armenian. The collective contribution of these researchers forms a robust foundation for further advancements in the field, highlighting the necessity for continuous exploration and innovation in speech recognition technology.

2.3 Progress in Morphology and Syntax

The development of NLP resources for underdocumented languages like Armenian has seen noteworthy progress in morphological and syntactic analysis, as highlighted by several key studies (Dolatian et al., 2022; Kindt and Kepeklian, 2022; Vidal-Gorène and Kindt, 2020; Vidal-Gorène et al., 2020; Ghukasyan and Avetisyan, 2021; Arkhangelskiy et al., 2012; Chiarcos et al., 2018).

Morphological Analysis: A wide spectrum of tools has been developed, including a morphological transducer by Dolatian et al. (2022) for Western Armenian, and a reusable RNN model by Vidal-Gorène et al. (2020) for various Armenian dialects. The latter indicates the possibility of rapid corpus processing in languages with limited NLP resources.

Lexical Analysis: Kindt and Kepeklian (2022) successfully applied a hybrid method combining digital dictionaries and RNN methodologies for analyzing ancient Armenian text, underscoring the utility of traditional and machine learning techniques.

Lemmatization and POS-Tagging: There have been advancements in the development of lemmatization and POS-tagging resources for Classical Armenian (Vidal-Gorène and Kindt, 2020), including the creation of a lightweight yet highly accurate lemmatizer (Ghukasyan and Avetisyan, 2021).

Corpus Creation: Corpus construction tools, like UniParser and the EANC platform, were utilized by Arkhangelskiy et al. (2012) to construct a morphologically annotated corpus for a minority language, demonstrating the potential for similar initiatives in other low-resource languages.

Extensions to Universal Morphologies: Chiarcos et al. (2018) aimed at enhancing the coverage of Universal Morphologies (UniMorph) for the languages of the Caucasus region, pointing towards the potential for further improvement and adaptability of existing tools.

In summary, there is promising progress in the morphological and syntactic analysis of Armenian, with studies indicating the potential of machine learning methodologies and the need for larger, more representative datasets. However, further research is needed to augment the adaptability of these tools across different Armenian dialects and

improve their accuracy.

2.4 Advancements in Semantics and Named Entity Recognition (NER)

Developments in semantics and Named Entity Recognition (NER) for Armenian have been noteworthy, as exhibited by the research conducted by Mkhitaryan and Madatyan (2022); Podolak and Zeinert (2020); Jain et al. (2019); Tambuscio and Andrews (2021); Vachagan and Tigran (2015); Ghukasyan et al. (2018).

Semantics: Exploring temporal expressions in Armenian fairy tales, Mkhitaryan and Madatyan (2022) identified shared schemas with English, thus guiding the future semantic analysis of Armenian texts. Additionally, Vachagan and Tigran (2015) developed algorithms for advanced sentence-level semantic analysis, addressing issues like homonyms.

Cross-Lingual NER: Multilingual transformer models were investigated by Podolak and Zeinert (2020) to understand the influence of language-specific features on performance. Meanwhile, Jain et al. (2019) leveraged machine translation to enhance NER in languages with sparsely annotated corpora, exhibiting superior performance for Armenian NER compared to a monolingual model.

Geolocation NER: Tambuscio and Andrews (2021) enhanced the detection of geographical entities from historical Armenian text by combining the output of multiple NER tools and utilizing geographical clustering, overcoming the limitations of traditional NER tools for older texts.

Resources for Armenian NER: Ghukasyan et al. (2018) made a significant contribution by providing a silver- and gold-standard dataset, setting baseline results for popular models, and sharing Armenian word embeddings, all of which can aid in the development of Armenian NER systems.

In summary, the progress made in Armenian semantics and NER is considerable. However, the potential for improvement is vast, with future directions including refining models, creating more representative datasets, leveraging machine translation for cross-lingual NER, and enhancing semantic analysis techniques. These advancements are essential for fully understanding and preserving the richness of the Armenian language.

2.5 Innovations in Document Analysis and Optical Character Recognition

Pioneering developments in document analysis and optical character recognition (OCR) have significantly benefited the preservation and accessibility of low-resource languages, such as Armenian. The works by Vidal-Gorène et al. (2021); Hovakimyan et al.; Islam and Dundua (2015); Tigranyan and Ghukasyan (2020) illustrate this trend:

Multilevel Annotation Platform: Vidal-Gorène et al. (2021) created an online, modular interface to annotate handwritten and printed documents. By developing an Armenian manuscript database, they showed the utility of such platforms in automating tasks and managing data. Their tool demonstrates the potential to enrich digital humanities resources, particularly for lesser-documented languages.

Text Recognition: Hovakimyan et al. introduced a novel two-layer neural network approach for Armenian text recognition. By outperforming single-layer networks, their work significantly advances the digitization of Armenian texts, expanding the possible applications of complex networks in text recognition.

Origin Tracing: Islam and Dundua (2015) employed supervised machine learning to identify the origins of the Georgian Gospel. Their classifier, based on lexical and classical information-theoretic features, distinguishes original from translated documents, underscoring the potential of machine learning in document analysis and multilingual data handling.

Error Reduction: The challenge of OCR errors in Armenian texts was addressed by Tigranyan and Ghukasyan (2020). Through a two-step method involving a multilayer perceptron and a convolutional neural network-based sequence transducer, they reduced the word error rate in OCR output, emphasizing the importance of post-processing in OCR tasks.

In summary, these studies signify considerable progress in document analysis and OCR for Armenian, revealing the power of innovative methods and collaborative platforms. However, further research is needed to continue enhancing these resources' quality, accuracy, and breadth, ensuring the preservation and accessibility of Armenian texts.

2.6 Emerging Trends in Armenian Text Similarity and Plagiarism Detection

Recent studies in text similarity and plagiarism detection have introduced promising methodologies, particularly for low-resource languages like Armenian. The works by Avetisyan et al. (2023);

Yeshilbashian et al. (2022); Ter-Hovhannisyan and Avetisyan (2022); Margarov et al. (2017) contribute to this growing field:

Cross-Lingual Plagiarism Detection: Avetisyan et al. (2023) introduced a cross-lingual plagiarism detection method leveraging multilingual thesauri and pre-trained BERT-based language models. The method's effectiveness across languages and its independence from machine translation or word sense disambiguation make it a fitting solution for less-resourced languages.

Stylometric Techniques: Yeshilbashian et al. (2022) used intrinsic stylometric techniques for Armenian text plagiarism detection. They implemented hierarchical clustering models, thereby indicating areas for improvement for longer documents and suggesting refinements for feature sets and parsers.

Cross-Lingual Sentence Alignment: The value of pre-trained Transformer-based language models for cross-lingual sentence alignment was underscored by Ter-Hovhannisyan and Avetisyan (2022). They found the XLM-RoBERTa model outperformed others across multiple languages, reinforcing the advantage of a single model trained on various languages in a multilingual context.

Multi-Layered Analysis: Margarov et al. (2017) introduced an Armenian text similarity analysis system. By conducting a multi-layered analysis and facilitating user interaction, their approach comprehensively detects plagiarism, including synonym replacement and translation-based plagiarism.

These studies highlight the potential of advanced language models, stylometric techniques, and user-interactive systems for enhancing text similarity and plagiarism detection in Armenian and other low-resource languages. They lay out directions for future research, including refining feature sets, devising efficient model training strategies, and creating more user-interactive systems.

2.7 Evolution of Language Modeling for Armenian Language

The field of language modeling has seen significant growth, with research focusing on creating custom models for the Armenian language and enhancing the cross-lingual capabilities of multilingual models.

Compact Language Modeling: In Karamyan and Karamyan (2022), a compact and efficient N-gram language model was developed specifically for Armenian. The researchers optimized the model's size and performance using pruning, quantization, and Byte Pair Encoding (BPE) techniques. This research resulted in a compact, high-performing language model for Armenian.

Cross-lingual Transfer: In an endeavor to enhance the cross-lingual transfer abilities of Multilingual BERT (mBERT), Kulshreshtha et al. (2020) managed to improve its performance. The researchers achieved this by aligning mBERT with cross-lingual signals, employing parallel corpora supervision, and fine-tuning the alignment. Their study provides valuable insights into strengthening language transfer in multilingual models.

Word Embedding Models: On a different track, Avetisyan and Ghukasyan (2019) tested existing Armenian word embedding models and introduced new ones. They employed intrinsic and extrinsic evaluation methods involving a variety of language tasks for comprehensive analysis. The researchers found that different models excelled at different tasks, highlighting the inadequacy of a "one-size-fits-all" approach in choosing an embedding model. They also provided new embeddings that outperformed existing ones in several tasks, contributing to more effective Armenian language processing tools.

In conclusion, efforts in Armenian language modeling are shaping the landscape of language-specific models, enhancing multilingual models' cross-lingual capabilities, and refining word embeddings. These research endeavors contribute to a richer understanding and processing of the Armenian language, driving progress toward more effective natural language processing tools.

3 Trends Over Time in Armenian NLP Research

Research on Armenian Natural Language Processing (NLP) has evolved significantly, showcasing notable trends in methods and techniques employed for addressing NLP tasks in the Armenian language.

Initially, Armenian NLP research focused on developing basic tools and methods, such as keyword identification and stemming, to handle specific tasks like plagiarism detection (Margarov et al., 2017). These early studies aimed to overcome the language's unique challenges and lay the founda-

tion for further advancements.

As research progressed, more advanced techniques and algorithms were adopted. Stylometric approaches emerged, enabling the detection of style changes and breaches in Armenian texts (Yeshilbashian et al., 2022). These findings highlighted the potential for more sophisticated language-specific tasks.

Recently, there has been a shift towards utilizing transformer-based models in Armenian NLP. Pretrained multilingual models like BERT and XLM-RoBERTa have shown promising results in tasks requiring a deeper understanding of text semantics and cross-lingual alignment (Avetisyan et al., 2023; Ter-Hovhannisyan and Avetisyan, 2022). However, the application of large language models in Armenian NLP remains relatively unexplored.

In conclusion, Armenian NLP research has transitioned from basic text processing techniques to embracing complex transformer-based models. While significant progress has been made, exploring large language models in Armenian NLP is still in its early stages. This trend emphasizes the need for further research, particularly in applying large language models, to address the unique challenges of the Armenian language and drive the field forward.

4 Identification of Gaps, Limitations, and Challenges in Current Research

Based on our review, the main challenges in Armenian Natural Language Processing (NLP) are as follows:

1. **Limited Availability of Large, Labeled Datasets**: Armenian NLP is considerably constrained by the scarcity of large and diverse labeled datasets (Khurshudyan et al., 2022). These datasets are essential for training and evaluating machine learning models, including large language models (LLMs). The lack of such resources limits the capacity to leverage the full potential of LLMs for Armenian NLP tasks.

2. **Shortage of Linguistic Resources and Tools**: The lack of comprehensive linguistic tools, such as parsers, lemmatizers, and tokenizers, poses significant challenges for Armenian NLP. This lack becomes particularly evident in tasks requiring complex linguistic analysis, like stylometry (Yeshilbashian et al., 2022) and plagiarism detection (Avetisyan et al., 2023). Additionally, the absence of resources like a robust Armenian WordNet counterpart constrains the development of applications such as semantic analysis and machine translation.

3. **Less Prevalent Use of Advanced Models**: While studies are leveraging advanced machine learning models such as BERT-based and transformer models for Armenian NLP tasks (Avetisyan et al., 2023; Ter-Hovhannisyan and Avetisyan, 2022), the overall application of such models, especially large language models (LLMs), is still limited. This restricts the understanding and potential benefits of using LLMs in the Armenian NLP.

4. **Handling Dialectal Variations**: Dialectal variations of Armenian (Weitenberg, 2002) can significantly impact the performance of language models if not accounted for during their development and training.

5. **Challenges in Adapting Armenian for NLP Applications**: Armenian, as a unique Indo-European language with a rich history, faces several linguistic challenges integral to its speakers and the broader linguistic landscape. One of the distinctive features of Armenian is its alphabetic script (Sanjian, 1996), which has its own set of characters and symbols, making it distinct from the Roman or Cyrillic scripts. While culturally significant, this uniqueness can pose difficulties regarding digital representation and compatibility with international standards. Adapting Armenian to modern technological applications often requires accurate font development and encoding to ensure proper rendering and functionality. Moreover, the Armenian language has undergone various historical shifts and linguistic influences due to its location at the crossroads of Eastern Europe and Western Asia. These influences, including Persian, Arabic, and Russian, have left their mark on Armenian vocabulary and grammar. Navigating this linguistic history and preserving the purity of the language can be a complex undertaking. Economic factors and government policies are also crucial in shaping the language's trajectory, impacting its promotion, preservation, and accessibility.

Furthermore, handling inflections, word order flexibility, and the rich morphological system of Armenian (Dum-Tragut, 2009) could present unique challenges in developing practical NLP tools and models. These issues require in-depth research to create NLP solutions that accurately capture and process the complexities of the Armenian language. Thus, the intersection of linguistic uniqueness, historical influences, and complex linguistic features necessitates a multidimensional approach to address the challenges of adapting Armenian for modern NLP applications.

6. **Low Commercial Interest**: Due to the limited population of Armenian speakers compared to other languages, there may be less commercial incentive to invest in Armenian NLP research and development. This can slow down progress in this field.

5 The utilization of LLMs in Armenian NLP tasks

The recent advancements in natural language processing (NLP) have been significantly driven by the advent and development of large language models (LLMs), such as BERT, GPT-3, and T5. However, despite their evident success, the adoption of LLMs in Armenian NLP tasks has been somewhat limited, as reflected in the current literature (Avetisyan et al., 2023; Yeshilbashian et al., 2022; Ter-Hovhannisyan and Avetisyan, 2022; Margarov et al., 2017).

Only a handful of studies in Armenian NLP have explored the use of Large Language Models (LLMs). For instance, Avetisyan et al. (2023) employed a BERT-based model for cross-lingual plagiarism detection, showcasing its versatility. In a related effort, Ter-Hovhannisyan and Avetisyan (2022) utilized the transformer-based XLM-RoBERTa model for cross-lingual sentence alignment, highlighting its effectiveness in multilingual contexts. Additionally, Kulshreshtha et al. (2020) augmented the cross-lingual transfer abilities of Multilingual BERT (mBERT) to achieve superior performance in language transfer tasks.

Despite the proven efficacy of Large Language Models (LLMs) in various language processing tasks for well-resourced languages, their utilization in Armenian NLP remains limited. The prevailing approaches in Armenian NLP lean towards traditional methods, such as clustering, bag-of-words, and rule-based techniques (Yeshilbashian et al., 2022; Margarov et al., 2017). While these methods have shown relative success, the underutilization of LLMs highlights a research gap and the potential to enhance the performance of Armenian NLP systems.

Given the transformative effect of LLMs on NLP tasks in other languages, the underutilization of LLMs in Armenian NLP calls for future research. Exploring the reasons behind this limited application and identifying potential solutions could greatly benefit Armenian NLP. Therefore, future studies on Armenian NLP should consider the following directions:

1. **Expanding Data Resources**: One of the main challenges in Armenian NLP is the shortage of high-quality and diverse language resources. Future work should prioritize developing and augmenting Armenian language corpora, including web scraping of Armenian text, transcription of spoken language, and translation of existing datasets from other languages. Efforts should be made to ensure these datasets span multiple domains and genres, such as news, academic literature, fiction, and social media posts (Avetisyan et al., 2023).

2. **Fine-tuning of Pre-trained Models**: As demonstrated in studies (Ghukasyan et al., 2018; Avetisyan et al., 2023; Ter-Hovhannisyan and Avetisyan, 2022), fine-tuning pre-trained language models on Armenian-specific tasks can yield impressive results. However, this approach could be extended to other NLP tasks that have not been extensively explored in the Armenian context, such as sentiment analysis, topic modeling, text summarization, and dialogue systems.

3. **Investigating Language-specific Challenges**: Armenian, like any other language, possesses unique linguistic characteristics that can pose challenges for NLP. These characteristics may include complex morphology, dialects, or historical language changes. Understanding these features and how they affect the performance of LLMs is crucial. Future work could explore these challenges and devise strategies to mitigate their effects (Yeshilbashian et al., 2022).

4. **Developing Armenian NLP Applications**: Once sufficient resources and fine-tuned models are in place, efforts should build practical applications for Armenian NLP. These applications could include Armenian sentiment analysis systems, named entity recognizers, chatbots, and machine translation systems (Margarov et al., 2017).

5. **Benchmarking and Evaluation**: Establishing benchmarks for various Armenian NLP tasks is crucial. These benchmarks will provide a clear direction for the research community and enable comparing different models and approaches.

6. **Developing an Armenian-specific Language Model**: While multilingual models have succeeded, creating a large pre-trained language model specifically for Armenian could lead to even better performance. This would require collecting and preprocessing a large Armenian corpus, which could benefit the wider Armenian NLP community.

7. **Cross-lingual Transfer Learning**: While Avetisyan et al. (2023) have successfully used multilingual models, there is a need for more research on techniques for transferring learning from high-resource languages to Armenian. This can help achieve higher accuracy in NLP tasks without requiring extensive Armenian-specific datasets.

8. **Exploration of Smaller Models**: Large language models require significant computational resources due to their size. Exploring the application and performance of smaller models or efficient training methods could enable the broader use of language models in Armenian NLP.

The potential benefits of these efforts are wide-reaching. From a research perspective, they can improve performance across various NLP tasks, deepening our understanding of Armenian language processing. For users, these advancements could lead to more accurate and helpful language technology applications, including better machine translation systems, more effective information retrieval tools, more engaging dialogue systems, and more robust text analysis tools.

In conclusion, by leveraging the potential of LLMs and addressing the unique challenges of Armenian NLP, researchers and practitioners can pave the way for significant advancements in language technology for Armenian. The proposed directions and future work outlined in this paper aim to stimulate further research and collaboration in the field, ultimately contributing to the growth and development of Armenian NLP.

6 Conclusion

In conclusion, this paper has taken significant strides in illuminating the state of Natural Language Processing (NLP) research in the Armenian language context, a low-resource language that has been relatively under-explored in the computational linguistics landscape.

Our analysis highlights a substantial gap in deploying Large Language Models (LLMs) in Armenian NLP tasks, despite their proven effectiveness across various NLP tasks for well-resourced languages. This finding underscores the potential of harnessing the power of LLMs to boost Armenian language processing tasks. However, it simultaneously reveals formidable challenges, such as the scarcity of training data, substantial computational resource requirements, and the peculiarities of the Armenian language.

With the rapid advancements in NLP and the growing focus on low-resource languages, our study articulates the imperative and the potential for further exploration of Armenian NLP, armed with the power of LLMs. Our work catalyzes such exploration, setting the stage for what we anticipate will be a future filled with robust development in this area. As part of our contributions, we have also suggested several directions for future work, including training LLMs specifically on Armenian data or fine-tuning existing multilingual models.

Moreover, our paper makes a critical contribution by situating the discourse on Armenian NLP within the broader NLP landscape and emphasizing the need for the NLP community to give equal attention to low-resource languages. This can lead to development of more sophisticated NLP tools for these languages and contribute to a more comprehensive understanding of language diversity—an aspect crucial for the holistic advancement of NLP.

We would like to encourage our fellow researchers and practitioners in NLP to explore low-resource languages, such as Armenian. Investigat-

ing such languages holds immense potential for revealing novel approaches and techniques in NLP that can serve a broader range of languages and cultures. This enriches the diversity in the NLP field and presents a unique opportunity to refine and innovate our methodologies.

Engaging with low-resource languages can also significantly reduce the digital divide and promote diversity, ensuring all languages find representation in the digital space. So, let us collectively embrace the challenges of low-resource languages and uncover the vast, untapped potential they offer to advance NLP.

Lastly, it is crucial to acknowledge that the development of Armenian NLP goes beyond technological advancements. It encompasses the preservation of language and cultural heritage. In today's digital world, languages lacking essential NLP resources face the risk of being marginalized. Hence, investing in Armenian NLP ensures the longevity and vitality of the Armenian language in the digital age.

Limitations

While providing a thorough review of Armenian NLP research, our study has several limitations that should be noted. Primarily, the scope of our review is confined to the set of studies and papers we have access to, which might not encompass the full breadth of relevant work in this area. Further, exploring Armenian as a low-resource language, with its associated challenges, is a broad and multifaceted issue. Although we strive to cover this topic comprehensively, there could be facets that were not fully captured in our discussion.

It is also essential to underscore that our review fundamentally relies on the accuracy and thoroughness of the original papers incorporated into our analysis. Consequently, any inaccuracies or misinterpretations within those works could inadvertently impact the findings of our review.

Moreover, while we endeavored to present a rich snapshot of the current state of Armenian NLP research, the inherently static nature of a review such as this cannot entirely keep pace with the rapid advancements within the field. As such, certain parts of our analysis could potentially become outdated in a relatively short period.

Finally, our paper primarily emphasizes using large language models in Armenian NLP. In focusing on this aspect, there is a chance that other potentially effective approaches, including more traditional linguistic methods, may not have been as extensively explored.

Our suggestions for future work in Armenian NLP represent a projection of potential research directions. Nonetheless, the feasibility and effectiveness of these suggestions would ultimately need to be empirically validated in subsequent research.

Ethics Statement

We recognize the importance of ethical considerations in research and the broader impact of our work. Throughout the process of conducting this literature review, we have taken into account relevant ethical aspects.

This paper is solely focused on analyzing and synthesizing existing literature without involving human subjects, data collection, or experimentation. As a result, it does not require specific ethical approvals or considerations regarding human subjects, privacy, or data protection.

However, we are dedicated to promoting ethical practices and responsible information use. We have exercised caution to appropriately cite and attribute all the sources used in this review, respecting the intellectual property rights of authors and organizations. We have aimed to accurately and objectively represent the content of the literature, ensuring that no biases or misinterpretations are introduced.

Additionally, we understand the significance of inclusivity and diversity within the field of NLP. In our review, we have deliberately covered a broad range of studies and perspectives to provide a comprehensive and balanced overview of the literature on Armenian NLP. We have consciously tried to accurately represent the Armenian language and its speakers, actively avoiding biases or stereotypes. Our methodologies and models have been developed with fairness and equity in mind, striving to minimize potential biases or discriminatory effects.

Furthermore, we acknowledge the potential societal implications of our work. Our research aims to positively contribute to the Armenian language community by supporting language preservation, linguistic research, and application development for Armenian speakers.

References

Timofey Arkhangelskiy, Oleg Belyaev, and Arseniy Vydrin. 2012. The creation of large-scale annotated corpora of minority languages using uniparser and

the eanc platform. In *Proceedings of COLING 2012: Posters*, pages 83–92.

Levon Aslanyan, Minoosh Heidari, Hasmik Sahakyan, Daryoush Alipour, Jorge Fernández, Angel Castellanos, Juan Castellanos, Leonid Hulianytsky, Anna Pavlenko, Mariam Haroutunian, et al. 2015. On string mining speech recognition.

Karen Avetisyan. 2022. Dialects identification of armenian language. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 8–12.

Karen Avetisyan and Tsolak Ghukasyan. 2019. Word embeddings for the armenian language: intrinsic and extrinsic evaluation. *arXiv preprint arXiv:1906.03134*.

Karen Avetisyan, Arthur Malajyan, and Tsolak Ghukasyan. 2023. A simple and effective method of cross-lingual plagiarism detection. *arXiv preprint arXiv:2304.01352*.

Varuzhan H Baghdasaryan. Armenian speech recognition system: Acoustic and language models.

Varuzhan H Baghdasaryan. 2022. Armspeech: Armenian spoken language corpus. *International Journal of Scientific Advances (IJSCIA)*, 3(3):454–459.

Samuel Chakmakjian and Ilaine Wang. 2022. Towards a unified asr system for the armenian standards. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 38–42.

Christian Chiarcos, Kathrin Donandt, Maxim Ionov, Monika Rind-Pawlowski, Hasmik Sargsian, Jesse Wichers Schreur, Frank Abromeit, and Christian Fäth. 2018. Universal morphologies for the caucasus region. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Karamyan Davit and Karamyan Tigran. 2022. A conformer based automated speech recognition for armenian language. *//Scientific Artsakh*, (2 (13)):224–229.

Hossep Dolatian, Daniel Swanson, and Jonathan Washington. 2022. A free/open-source morphological transducer for western armenian. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 1–7.

Jasmine Dum-Tragut. 2009. *Armenian: Modern eastern armenian*, volume 14. John Benjamins Publishing.

T Ghukasyan and K Avetisyan. 2021. Research and development of a deep learning-based lemmatizer for the armenian language. - , page 92.

Tsolak Ghukasyan, Garnik Davtyan, Karen Avetisyan, and Ivan Andrianov. 2018. pioner: Datasets and baselines for armenian named entity recognition. In *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61. IEEE.

Anna Hovakimyan, Narine Ispiryan, and Gevorg Narimanyan. Armenian texts recognition via neural networks.

Zahurul Islam and Natia Dundua. 2015. Finding the origin of a translated historical document. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 96–105.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Davit S Karamyan and Tigran S Karamyan. 2022. Compact n-gram language models for armenian. *Mathematical Problems of Computer Science*, 57:30–38.

Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 2022. Eastern armenian national corpus: State of the art and perspectives. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 28–37.

Bastien Kindt and Gabriel Kepeklian. 2022. Analyse automatique de l'ancien arménien. évaluation d'une méthode hybride «dictionnaire» et «réseau de neurones» sur un extrait de l'adversus haereses d'irénée de lyon. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 13–20.

Bastien Kindt and Emmanuel Van Elverdinghe. 2022. Describing language variation in the colophons of armenian manuscripts. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 21–27.

Saurabh Kulshreshtha, José Luis Redondo-García, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual bert: A comparative study. *arXiv preprint arXiv:2009.14304*.

Arthur Malajyan, Karen Avetisyan, and Tsolak Ghukasyan. 2020. Arpa: Armenian paraphrase detection corpus and models. In *2020 Ivannikov Memorial Workshop (IVMEM)*, pages 35–39. IEEE.

Gevorg Margarov, Gohar Tomeyan, and Maria João Varanda Pereira. 2017. Plagiarism detection system for armenian language. In *2017 Computer Science and Information Technologies (CSIT)*, pages 185–189. IEEE.

Karine Megerdoomian. 2009. Low-density language strategies for persian and armenian. In *Language Engineering for Lesser-Studied Languages*, pages 291–312. IOS Press.

Yelena Mkhitaryan and Lusine Madatyan. 2022. A spatial model of conceptualization of time: With special reference to english and armenian fairy tales. *International Journal of Language and Culture*.

Jowita Podolak and Philine Zeinert. 2020. Master thesis: Developing a cross-lingual named entity recognition model.

Jatinderkumar R Saini and Rajnish M Rakholia. 2016. On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. *Procedia Computer Science*, 89:313–319.

Avedis K Sanjian. 1996. The armenian alphabet. *The world's writing systems*, pages 356–357.

Sona Sargsyan and Torsten Rahne. 2021. Development of speech material for an armenian speech recognition threshold test. *Russian Open Medical Journal*, 10(3):321.

Max Silberztein. 2007. An alternative approach to tagging. In *Natural Language Processing and Information Systems: 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27-29, 2007. Proceedings 12*, pages 1–11. Springer.

Marcella Tambuscio and Tara Lee Andrews. 2021. Geolocation and named entity recognition in ancient texts: A case study about ghewond's armenian history. In *CHR*, pages 136–148.

Tatevik Ter-Hovhannisyan and Karen Avetisyan. 2022. Transformer-based multilingual language models in cross-lingual plagiarism detection. In *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 72–80. IEEE.

S Tigranyan and T Ghukasyan. 2020. Post-ocr correction of armenian texts using neural networks. *Vestnik" Scientific Journal of Russian-Armenian University.—2020*.

Vahradyan Vachagan and Apozyan Tigran. 2015. On meanings of words, sentences and texts interpreted for chess and literary eastern armenian.

Chahan Vidal-Gorène and Aliénor Decours-Perez. 2020. Languages resources for poorly endowed languages: The case study of classical armenian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3145–3152.

Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. 2021. A modular and automated annotation platform for handwritings: evaluation on under-resourced languages. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16*, pages 507–522. Springer.

Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 2020. Recycling and comparing morphological annotation models for armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101.

Chahan Vidal-Gorène and Bastien Kindt. 2020. Lemmatization and pos-tagging process by using joint learning approach. experimental results on classical armenian, old georgian, and syriac. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27.

Jos JS Weitenberg. 2002. Aspects of armenian dialectology. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 137:141–158.

Ye M Yeshilbashian, AA Asatryan, and Ts G Ghukasyan. 2022. Plagiarism detection in armenian texts using intrinsic stylometric analysis. *Programming and Computer Software*, 48(7):435–444.

A Literature Search and Keywords

The literature search was conducted in prominent academic databases such as the ACL Anthology, IEEE Xplore, Google Scholar, Semantic Scholar, ACM Digital Library, SpringerLink, and ScienceDirect. These databases were chosen due to their extensive coverage of relevant research papers in the field of Natural Language Processing (NLP).

To retrieve relevant literature, we employed a comprehensive set of keywords. These keywords were carefully selected to encompass various aspects of NLP research, including methodologies, techniques, applications, and specific subdomains within NLP. The chosen keywords aimed to capture a broad range of literature pertinent to our research objectives.

By combining general and specific keywords, we sought to cast a wide net to ensure the inclusivity of relevant publications. This approach enabled us to explore diverse perspectives and comprehensively understand the current state of research in NLP.

The selected keywords were rigorously reviewed and refined iteratively to optimize their effectiveness in retrieving relevant literature. The process involved considering synonymous terms, related concepts, and variations in terminology commonly used in the NLP community. This ensured that our search strategy was comprehensive and accounted for different linguistic expressions and terminologies used across the literature.

The selected keywords are as follows: "Armenian language" AND "Natural Language Pro-

cessing" "Armenian" AND "NLP" "Armenian language" AND "Machine Learning" "Large Language Models" AND "Armenian" "Low-resource" AND "NLP" AND "Armenian" "Armenian language" AND "deep learning" "Transfer learning" AND "Armenian" "Armenian" AND "BERT" "Armenian" AND "Transformer models" "Armenian language" AND "AI" "Neural Networks" AND "Armenian language" "Armenian" AND "Language Models" "Armenian" AND "Text classification" "Armenian" AND "Named Entity Recognition" "Armenian" AND "Sentiment Analysis" "Armenian" AND "Speech Recognition" "Armenian" AND "Language Generation" "Armenian" AND "Text-to-Speech".