The First Parallel Corpus and Neural Machine Translation Model of Western Armenian and English

Ari Nubar Boyacıoğlu, Jan Niehues
Karlsruhe Institute of Technology, Karlsruhe, Germany
me@arinubar.com, jan.niehues@kit.edu

Abstract

Western Armenian is a low-resource language spoken by the Armenian Diaspora residing in various places of the world. Although having content on the internet as well as a relatively rich literary heritage for a minority language, there is no data for the machine translation task and only a very limited amount of labeled data for other NLP tasks. In this work, we build the first machine translation system between Western Armenian and English. We explore different techniques for data collection and evaluate their impact in this very low-resource scenario. Then, we build the machine translation system while focusing on the possibilities of performing knowledge transfer from Eastern Armenian. The system is finetuned with the data collected for the first Western Armenian-English parallel corpus, which contains a total of approximately 147k sentence pairs, whose shareable part of 52k examples was made open-source. The best system through the experiments performs with a BLEU score of 29.8 while translating into English and 17 into Western Armenian.

Keywords: Western Armenian, parallel corpus, machine translation

1. Introduction

The advancements in the fields of Deep Learning and Natural Language Processing (NLP) have made a significant impact on the daily lives of people, in the global markets as well as shifted the trajectory of research. The introduction of the internet has made the world a little bit smaller by bringing communities together in a single platform. Perhaps the biggest remaining hurdle in this process, the language barrier, was finally eliminated with the inclusion of machine translation tools.

The dependence of deep neural models on large amounts of data has brought an important phenomenon that became an important topic in NLP research: Not all languages enjoyed the advancements in NLP equally, but only the ones that have a proper presence on the internet and that have content which is easily usable and can be converted to training material for neural models fulfilling a specific NLP task. This effectively resulted in a divide between high and low-resource languages, where, as the names suggest, languages have high or low amounts of training material and therefore do not acquire the same support in the research and the same representation in the end-products of NLP. This phenomenon was observed by (Joshi et al., 2021), stating the research mainly focuses on a handful of (related) languages where the vast majority of linguistic phenomena are ignored. Low-resource languages and establishing a proper diversity of language technologies is a great challenge and a highly active research area. Giving the same treatment to every language not only helps build stronger connections between various communities of the world but also preserves and adds resistance to the process of language extinction. (Rehm and Way, 2023)

In this work, we investigate the rather neglected variant of Modern Armenian: Western Armenian (WA), which is mainly spoken by the Armenian Diaspora residing in the Americas, Europe, the Middle East, and Australia and is classified as an endangered language by UNESCO (2010). It has an active community producing various content on the internet, as well as a literary heritage coming from the 19th century, yet it lacks the datasets curated for building Neural Machine Translation (NMT) and other NLP systems. Our work focuses on building the first NMT system that supports WA and creating its first parallel corpus. We conduct an extensive search on the internet and the printed media for finding suitable candidates for WA resources while aiming to have a fair range of domains. The collected data was utilized in different experiments to assess and evaluate the impact of the translation quality using automatic metrics. Additionally, since WA resources are currently limited and its cognate language Eastern Armenian (EA) has relatively more resources in terms of available training data and shares a fair portion of similarities with WA, we investigate the possibility of EA knowledge transfer within the pre-trained models or through additional finetuning. The part of our corpus, which does not get subjected to any copyright is available online[1] and contains approximately 52k sentence pairs.

[1] https://github.com/AriNubar/hyw-en-parallel-corpus

2. Armenian and its Modern Variants

Armenian is a language belonging to the Indo-European language family which is written with the Armenian alphabet, consisting of 38 letters. It is an inflected language, with no gender and mainly adopts the (S)VO and (S)OV sentence structure. The modern variants of Armenian emerged from Classical and Middle Armenian in the 18th century by adopting themes of the common folk in its literature, as well as the dialects of the then major centers of Armenian communities of the Ottoman and the Russian Empire: Bolis (Constantinople) dialect for Modern Western Armenian and Tiflis (Tbilisi) dialect for Modern Eastern Armenian, while the latter has adopted Yerevan dialect subsequently. Both variants have shown individual development paths due to interactions with different languages, however, they stayed mutually intelligible (Campbell, 2003; Donabedian-Demopoulos, 2018), although speakers of one variant may need to adapt themselves while listening to the other variant or reading it, since there are differences in grammar, intonation, vocabulary and orthography.

Both variants have been classified as separate languages by the SIL ISO 639-3 Registration Authority (2017), whose report states that both variants' "linguistic distance is not great, but having developed distinct vocabularies and literature is the evidence for the emergence of two languages." The languages are represented thus with separate codes of `hyw` for WA and `hye` for EA.

Modern Eastern Armenian is the official language of the Republic of Armenia and is mainly spoken in the countries of the Eastern Bloc as well as by the individuals who emigrated from the countries of the Soviet Union to the United States and various countries of Europe. Modern Western Armenian is a diasporic language, spoken currently by the descendants of individuals who have survived the Armenian Genocide in the early 20th century and emigrated to many countries over the world. Due to its diasporic nature, the language suffers from the problems of being a minority language: no official representation, difficulties in making a modernized curriculum, having to rely on voluntary efforts, limited representation, and slow adaptation to the modern environments such as the internet, some of its speakers deliberately choosing not to pass down knowledge to further generations in order to have a better integration process to the host country; effectively showing symptoms of a dying language.

The phrase "the Armenian language" usually refers to the Eastern variant in practice. For WA-speaking communities, this is one of the major struggles, and many personal and organizational projects are dedicated to resisting and eliminating the threat of language death with campaigns to raise awareness of the issue, international programs to train educators, projects to extend WA's usage other than homes, to modernize and introduce the language to the rest of the world. Ethnologue (2023) states that WA has 1.6 million speakers worldwide, whereas EA is spoken by 3.7 million people. Although Armenian is recognized as a minority language in various European countries that have signed the European Charter for Regional or Minority Languages (Council of Europe, 1992), any of its modern variants has been mentioned in the recently published book of the European Language Equality project (Rehm and Way, 2023), which aims to establish political equality for all languages in Europe.

2.1. Western Armenian

Previous works (Goyal et al., 2022; Heffernan et al., 2022; Izbicki, 2022; Kann et al., 2020; Yu et al., 2020) mention (Eastern) Armenian as a low-resource language, but they lack the distinction between the Eastern and Western variants, referring exclusively to EA.

Nevertheless, there are some works from the late 2000s and more recently in the late 2010s-early 2020s about WA data collection/corpus building as well as some NLP models. The first annotated dataset of WA was created by Donabedian-Demopoulos and Boyacioglu (2007) using NooJ (Silberztein, 2005), a software for formalizing natural language and annotating textual data. They use the works of WA authors of the late 19th century as the corpus, which is partially available on the official NooJ website (NooJ, 2023). Additionally, Khachatryan (2011; 2012) uses NooJ for annotating and creating a formal grammar on WA nouns using an individual WA printed press corpus. More recently, as a part of Universal Dependencies treebank project (de Marneffe et al., 2021), Yavrumyan (2023) releases WA-ArmTDP, a syntactically annotated corpus in treebank format. The corpus contains a total of ca. 120k tokens over 6656 sentences. Boyacioglu and Dolatian (2020) release a list of verb conjugation paradigms along with a sample list of 3000 verbs. The paradigms are implemented in an open-source rule-based morphological transducer created by Dolatian et al. (2022), which is suitable to the Apertium environment (Forcada et al., 2011). Dolatian et al. (2022) share the corpus, which is used during the implementation and testing of the transducer and contains scraped texts from the WA Bible, Wikipedia, and media. Building a large syntactically and semantically annotated corpus for WA is one of the main parts of the ongoing "Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus" project by INALCO, which was initiated in April 2021. Finally,

the search for existing MT resources for WA did not return any results during the course of this work.

The last couple of years have seen some activity in NLP research for WA: Avetisyan (2022) compares statistical and neural models for disambiguating the modern Armenian variants as well as Classical Armenian. Neural models achieve a 98% accuracy across all languages. Vidal-Gorène et al. (2020) create neural-based systems for the lemmatization and PoS-tagging tasks and compare their performance with rule-based systems. The EA-trained neural system outperforms on both EA and WA test sets. The authors state that EA-trained neural models could be used as a starting point in order to process WA unannotated texts (Vidal-Gorène et al., 2020). For speech recognition, Chakmakjian and Wang (2022) undertake a surveying work investigating the requirements, available data, and challenges for building a unified Western and Eastern Armenian speech recognizer, while The National Center of Communication and Artificial Intelligence Technologies (2022) aims to build Western and Eastern Armenian speech corpora by providing a platform where WA and EA speech data is collected via crowdsourcing.

WA is currently not included in any closed-source machine translation service, except the website provided by ISMA[2], however, it seems to be incomplete and to have built on word/phrase-based generation. On their website, there is no documentation available about the implementation.

3. Low Resource Machine Translation

It is estimated that there are over 7000 languages spoken in the world (Ethnologue, 2023), however not all languages in the world are supported in today's NLP models developed by both research and industry. In the case of NMT, the state-of-the-art models reach human-level translation quality for some language pairs (Popel et al., 2020;Toral et al., 2018). This became possible from the advancement in deep learning techniques which are dependent on large amounts of parallel data with a scale of tens or hundreds of millions. Such amount of data is available only for a select few languages, typically paired with English, Chinese, Arabic, and other European languages. The remaining languages are called low-resource languages, which have limited amounts of data and when it comes to parallel data finding such corpora becomes an impossible task. It is often the case for a parallel corpus of a low-resource language being too noisy and covering a very specific corpus, usually including only the translations of religious texts. Low-resource languages suffer not only from the limited amounts of data but also from the lack of tools for collecting data, including language identifiers, OCR, bitext miners, etc.

There has been lately a trend to focus on non-English NMT, which mainly focuses on low-resource language pairs. Currently, the research in low-resource NMT focuses on implementing techniques to collect and prepare mono- and multilingual data as well as utilizing the knowledge about other (high-resource) languages for a low-resource language. For a more detailed overview, please refer to the survey works about low-resource MT (Ranathunga et al., 2023; Haddow et al., 2022; Wang et al., 2021).

Based on the classes theorized by Joshi et al. (2021), and on our estimation, WA belongs to Class 1, in which the languages have some unlabeled data online, and with some initiative, they may get better support from researchers. WA fits into this class because it has its own Wikipedia with 11627 articles as of February 2024 , as well as a multitude of news and other organizational websites; a fair amount of resources of WA texts yet not processed for NLP. WA has also the fortune to have a very close language: EA, which is often included in multilingual NLP systems; creating an opportunity to perform knowledge transfer to WA, although there is no previous work investigating this.

4. Data Collection

Before building the parallel corpus, the pairing language must be chosen. English is the most suitable choice for the first parallel corpus of WA since it is the lingua franca of the modern world, with which the research is mainly conducted; as well as it allows the parallel corpus and the translation model trained on it to reach the widest international audience possible. However, these languages have a relatively novel contact with each other, mainly because of the internet; meanwhile, languages like French, Turkish, and Arabic have had more interactions with WA throughout history. We plan on building parallel corpora with these languages in future works.

Originally, the search for parallel texts considered only online documents, however after a preliminary search, it has been decided that the online texts covered a relatively narrow range of domains, consisting mainly of religious and news domains. Thus, in order to extend the domain variety, the search was extended to consider printed media as well. This has also brought the opportunity to benefit from the old books which have become copyright-free.

The search for online resources started with the Wikipedia article "Armenian Newspapers"[3], where

[2] http://translator.am/en/index.html

[3] https://en.wikipedia.org/wiki/Armenian_newspapers

a list of active newspapers in Armenia and the rest of the world is presented. From there, a forward search was conducted through the links shared in the "Partners" section of each website. This led to a couple of organizational websites. After collecting a considerable amount of candidates, each of them was manually inspected so that they fulfill the constraints: 1) The multilingual material (if any) must contain English and WA parallel content. 2) The bilingual material must be direct translations of each other, allowing the least amount of alignment work. This was ensured by rigorously comparing bilingual material sentence by sentence. The oversights were planned to be reinspected in the manual correction step of the data preparation pipeline. 3) The bilingual documents must have a direct reference to each other (e.g. URL), eliminating the document alignment step. Additionally, WA and English Wikipedia were added to resources, even though they violate the second constraint. However, Wikipedia's wide domain coverage and popularity in many MT research works make it a prime resource.

For printed documents, the search was conducted in online and physical libraries in Germany, Turkey, and Armenia. We found out that finding the English translations of WA works was quite hard in the public libraries of said countries, therefore the search continued the other way around: finding WA translations of foreign authors. This has yielded better results since the libraries that have a collection of WA literature often include translations. Then, another search was conducted to determine whether the selected WA books had a digital version in an online library like the National Library of Armenia or required individual scanning and whether the English counterpart was included in open-source repositories like Project Gutenberg[4].

The National Library of Armenia provides a great share of their collection online[5], which serves as an invaluable resource for WA literature and printed media. The pieces in their collection are not labeled as WA or EA, at least in the online repository, so for the unfamiliar, it might be quite hard to disambiguate these languages. As a tip, one can make an advanced search by giving the place of publishing as a prompt. Typing major centers of WA-speaking communities (e.g. Պոլիս [Istanbul], Պէյրութ [Beirut], Փարիզ [Paris], Պոսթոն [Boston], Նիւ Եորք [New York] and Ֆրէզնօ [Fresno]; or countries like Թուրքիա [Turkey], Լիբանան [Lebanon], Ֆրանսա [France], and Միացեալ Նահանգներ [United States]) will result almost exclusively in WA books. Another important point about the books shared in online collection is that they are not fully digitized, but provided as scans; requiring an additional OCR step in the data preparation process.

[4] https://www.gutenberg.org/

[5] https://haygirk.nla.am/cgi-bin/koha/opac-main.pl

4.1. Resources

The research has resulted in various online and printed resources that make up the first WA-English parallel corpus. An overview of the statistics of the corpus can be found on Table 1 along with covered domains. Several datasets have been marked with a (*) both on Table 1 as well as on the titles of the following subsections, where each subset is briefly introduced. A starred dataset indicates that it is not redistributable and therefore excluded from the online repository.

4.1.1. Armeno-American Letter Writer (AALW)

Written by Haroutioun Hovannes Chakmakjian and published in 1914, it is a textbook case of a parallel corpus, as the left-hand side pages of this book are in WA and the right-hand side pages in English. The book is a collection of exemplary letters for various situations to teach how to write such letters, providing a unique domain of formal and informal correspondences as well as a rich selection of vocabulary.

4.1.2. The Bible

The Bible is often included in multilingual parallel datasets not only because it is written in many languages but it is quite trivial to align thanks to the verse numbers. The religious domain that the Bible covers, while limited, captures many personal and geographical names.

4.1.3. Gulbenkian Armenian Communities Newsletter (*)

Calouste Gulbenkian Foundation is a non-profit foundation that promotes and supports various art, science, and educational projects. It is currently regarded as the de facto language regulator of WA (Borjian, 2017) and organizes specialized projects for the preservation and development of the WA language. The dataset contains many modern words for technological concepts and a wide selection of Armenian names along with their English transliterations.

4.1.4. Hamazkayin Newsletter and Biographies

Hamazkayin Armenian Educational and Cultural Society is a major organization with multiple seats across the Armenian Diaspora. Hamazkayin organizes and supports many cultural events, such as exhibitions, festivals, seminars, book signings, etc. The Hamazkayin dataset was prepared from the news articles reporting the events hosted or sponsored by Hamazkayin as well as reviews about

many WA books and films. The dataset also includes biographies of individuals who have had an impact on the Armenian Diaspora, which are also contained on their website. Additionally, the names of countries and cities are very prominent in this dataset.

4.1.5. Hayern Aysor

Hayern Aysor (Armenians Today) is a news website established by the Diaspora Department of the "Center for Public Relations and Information" of the Republic of Armenia Prime Minister Office. It covers news from Armenia and the Armenian Diaspora along with official statements from the Armenian government, providing a unique domain. However, upon inspection, some WA articles seem that they were "modified" from EA, rather than being translated. This results in a unique mixed style of EA and WA.

4.1.6. Houshamadyan (*)

Houshamadyan is a project by a non-profit association in Berlin dedicated to preserving and showcasing the everyday life of the Armenian communities within various cities and the countryside of the Ottoman Empire. There are articles about local characteristics, education, economy, literature, traditions, clothing styles, and recipes of local dishes in WA, English, and Turkish. The dataset contains also a considerable amount of image captions.

4.1.7. The Watchtower Magazine of Jehovah's Witnesses (*)

This is another massively translated body of media that has included WA for many years. It includes articles about not only the Bible's prophecies but also perspectives on some contemporary topics like internet usage as well as personal stories, rendering it a multidomain resource. It includes personal names from many cultures along with their WA transliterations.

4.1.8. The Voice of Conscience (VoC) (*)

Written by the influential writer and politician of the late 19th century Krikor Zohrab, the book is a collection of short fictional stories in a realist manner. The book itself and its translation focus on maintaining a certain aesthetic which makes this dataset stylistically completely different from the other datasets within the corpus with its longer, descriptive sentences and usage of many stylistic devices.

4.1.9. WA Wikipedia

In NLP research, texts from Wikipedia articles are among the most commonly used data, due to their open-sourced nature and wide-range topic coverage. As resources of WA-English parallel texts are not plenty, we wanted to utilize Wikipedia because it includes unique topics and vocabulary, mainly originating from the domains of popular culture and science.

4.1.10. WA Monolingual Dataset

In low-resource MT, monolingual texts are often utilized to compensate for the scarceness of parallel texts. Using techniques like backtranslation, synthetic parallel datasets from monolingual datasets can be created. To investigate the effect of synthetic datasets, we collect an additional set of monolingual data from WA news websites: Jamanak, Agos, Aztag, and Arevelk.

Dataset Name	Domain	# Sent. Pairs	# WA Tok.	# EN Tok.
AALW	Correspondences (Formal & Informal)	2,135	31,225	38,658
Bible	Religious Texts	30,604	540,655	735,441
Gulbenkian (*)	News, Technology	598	10,680	13,453
Hamazkayin	News, Culture, Art, Literature, Education, Biographies	10,739	215,591	262,092
Hayern Aysor	News, Governmental, Official	5,422	92,920	115,139
Houshamadyan (*)	Sociology, Culture, Education, Food Recipes, Captions, Personal Stories	38,267	501,905	602,342
Watchtower (*)	Religion, Culture, Personal Stories, Philosophy	54,323	677,828	801,137
VoC (*)	Literature, Fictional Stories	889	32,331	37,636
hyw-Wikipedia	Biographies, Art, Science, Education, Literature, Geography, History, Popular Culture	3,979	76,156	100,293
HYW-Mono	News, Literature, Philosophy, Religion, Sports	1,437,035	26,056,315	31,890,452
	TOTAL Parallel Corpus	**146,956**	**2,179,291**	**2,706,191**
	TOTAL Open-Source	**52,879**	**956,547**	**1,251,623**
	TOTAL	**1,583,991**	**28,235,606**	**34,596,643**

Table 1: Datasets within the parallel corpus

4.2. Data Preparation Pipeline

Figure 1: Overview of Pipeline

Figure 1 illustrates the data preparation pipeline for the resources mentioned in the previous section, whose individual steps we describe below.

4.2.1. Collect & Shape

To digitize the printed documents, Tesseract OCR Engine (Ooms, 2023) was used. The engine's WA output however contains too many mistakes, which is probably caused by the engine's EA dictionary in the linguistic module. Although EA and WA share a substantial amount of vocabulary, they use different orthographies. After collection, the mistakes made by OCR were manually corrected.

For each website, a separate scraper script was written to collect documents on that website. There is no document alignment performed since the resources were chosen to contain bilingual documents that directly refer to each other.

Both types of collected documents are reshaped into lists of single sentences. To identify sentence boundaries automatically, the NLTK (Bird et al.,

2009) library was used for the English side which has a neural approach, and for the WA side, the rule-based pySBD (Sadvilkar and Neumann, 2020) library was used as NLTK lacked the support for either modern Armenian variant. The rules within the library were extended in the Armenian module to contain the ellipsis (...); additionally the colon (:) was added along with the Armenian sentence boundary character (։) as both are commonly used in practice as they look alike.

4.2.2. Automatic Alignment

Wikipedia articles can exist on the same topic across different languages, yet they are not always direct translations of each other. Often, they are referred to as comparable texts. Therefore, bitext mining was required in order to establish which Wikipedia articles were considered aligned translations. As WA currently does not have a language nor an MT model, we employ a method where each WA sentence within a document is translated into English using a couple of known machine translation services in the industry as if they were EA.[6] Each translated WA sentence is then compared with all English-side sentences for similarity, using NLTK's similarity score. The highest-scoring sentence that exceeds the score of 0.95 was chosen to be the counterpart for the WA sentence. This threshold was chosen after a qualitative investigation of the highest-scoring pairs as being actually the translations of each other.

4.2.3. Filtering

Any sentence pair containing emojis, URLs, or a long sequence of digits on either side is removed since these mainly bring noise instead of valuable information.

4.2.4. Manual Correction and Alignment

Sentence pairs from each document were compared and inspected line-by-line to make sure that they were direct translations. There were four major outcomes: 1) The pairs are complete direct translations of each other; 2) The pairs are direct translations of each other however there is additional information on either side; 3) The pairs are direct translations however they are spanned over a couple of sentence pairs (m-to-n alignment); 4) The pairs are not direct translations. Case 1 results in direct acceptance without any additional editing. In case 2, any additional information from either side is removed and afterwards, if the fluency of the sentence is not disrupted, the sentence is accepted. In case 3, aligning sentences were appended to each other to be contained in a single line. In other words, a single line contains multiple sentences for this example. Examples aligned to case 4 are eliminated.

[6]This is a common technique used in the WA-speaking community for translating WA into English. Although it is not documented, the translations are regarded as adequate enough to contain general information.

4.2.5. Final Filtering and Combination

Since the restructuring from the last step can introduce an imbalance of length for sentence pairs, another filtering step based on the sentence lengths was performed. Upon qualitatively inspecting the imbalanced sentences with various threshold values for length ratios, the value of 0.5 for either side was chosen. After eliminating unfulfilling pairs, all documents collected from a resource were combined into a single file, which is called a subset. Each subset is subdivided into train and test sets. The sizes of the train and set sets for a dataset were determined by the number of sentence pairs within that dataset. If the total amount of sentence pairs exceeds 4,000, then randomly sampled 2,000 non-repeating sentence pairs were included exclusively in the test set; if not, only 10% of the total amount of sentence pairs of the dataset was included.

5. Evaluation

With the help of experiments, we want to investigate several questions regarding WA machine translation. First, we focus on the usefulness of EA knowledge while performing WA translation. We investigate this in two scenarios: The zero-resource setting, where no WA data is available, and the low-resource setting when only small amounts of WA data are available. Previous works have shown adapting an NMT that was trained on a high-resource language was beneficial for improving the translation quality of a low-resource language in both directions as well as in both zero-resource (Ko et al., 2021) and low-resource (Maimaiti et al., 2019) settings.

Additionally, previous works have shown that the overlap of the domains within the training and test set plays a major role in obtaining high-quality translations, both in supervised and unsupervised settings (Liu et al., 2021; Kim et al., 2020; Marchisio et al., 2020; Siddhant et al., 2022). Domain adaptation of NMT models is a whole topic on its own with a plethora of works (Chu and Wang, 2018), however as of our knowledge there is no other work that compares the importance of (mis-)matching language information with the importance of (mis-)matching domain information within the train and test sets simultaneously. Therefore we conduct a second experiment where we train models with sin-

gular datasets that are contained in both EA- and WA-English parallel corpora.

5.1. Experiment Setup

As a baseline model, we choose the model "No Language Left Behind" of Team-NLLB et al. (2022), which is capable of translating between more than 200 languages, including EA, and has SOTA translation performance for many low-resource languages. This is done on the smallest version, NLLB-200-600M-Distilled, because of the limited amount of computational resources.

We then created different models by fine-tuning this model on the different data sets. Each finetuning session uses standard parameters and lasts for 5 epochs.

For additional EA-English parallel data we used the data shared in OPUS which are utilized in the models named NLLB + EA and NLLB + EA + WA and partially in NLLB + EA-Bible/Wiki (please refer to Table 2). For WA data, we use the data described in section 5.

Additionally, we utilize 3 synthetic datasets whose English sides are generated by models that are trained with genuine EA- and WA-English examples. We finetune individual models both with only synthetic datasets as well as a combination of authentic and synthetic examples. With these datasets, we aim to investigate: 1) what level of WA translation quality can be achieved with only EA-trained models and WA monolingual data; 2) whether including synthetic data along with genuine parallel examples improves the translation quality, as Poncelas et al. (2018) suggest that this is the case when the composition of synthetic and genuine data has a balance that is not tipped too far in favor of synthetic examples.

Finally, we make a doubly finetuned model, which is first trained with EA-English examples and then in an individual session with WA-English parallel examples. This model is an explicit representation of the utilization and transfer of EA knowledge.

The models in the first experiment are evaluated on the WA-test set. This set is the combined version of each test subset in the WA-English parallel corpus, as explained in data preparation pipeline. Synthetic datasets are not included in the test sets.

For an in-depth analysis, we focus on the effect of the matching domain against the matching language in training data. For this, we create specialized training and test sets that originate from the subsets found in both EA and WA parallel corpora and cover the same domain, i.e. the Bible and Wikipedia. We train 4 models for each language-subset combination and evaluate them on the WA Bible test set. We did not use Wikipedia, because it covers a wide range of domains which is not necessarily shared by the WA and EA counterparts and therefore can still bring domain mismatch.

For the names of the models in both experiments, please refer to Table 2.

Name	Description
Exp. 1: General Performance on Zero- and Low-Resource Settings	
NLLB	Baseline model with no additional finetuning.
+ EA	Finetuned with EA parallel examples.
+ WA	Finetuned with WA parallel examples.
+ EA + WA	Finetuned with EA parallel examples first, then separately with WA parallel examples.
+ sWA-mono$_{\text{NLLB + EA}}$	Finetuned with synthetic parallel examples, whose WA side is the monolingual dataset and English side is generated by NLLB + EA.
+ sWA$_{\text{NLLB + EA}}$	Finetuned with synthetic parallel examples, whose WA side is from the WA parallel dataset and English side is generated by NLLB + EA.
+ {WA, sWA-mono$_{\text{NLLB + WA}}$}	Finetuned with a balanced training data composition of genuine parallel WA examples and synthetic parallel examples whose WA side is the monolingual dataset and English side is generated by + WA.
Exp. 2: Domain vs. Language	
+ WA-Bible	Finetuned with WA Bible.
+ WA-Wiki	Finetuned with WA Wikipedia.
+ EA-Bible	Finetuned with EA Bible.
+ EA-Wiki	Finetuned with EA Wikipedia.

Table 2: Names of models in the experiments with their description.

We evaluate our results in each experiment using the automatic evaluation metrics of chrF3 (Popović, 2015) and BLEU (Papineni et al., 2002). Although BLEU is the most widely used automatic metric for MT tasks, it has received some criticism over the years (Stent et al., 2005; Callison-Burch et al., 2006; Ananthakrishnan et al., 2007; Smith et al., 2016). Since WA is an inflected language with a fair share of suffixes, BLEU becomes too strict of a metric. Therefore we include also the chrF3 score since its character-based scoring rewards partial matches.

5.2. Transfer Between Languages

Evaluated on: WA-test				
Direction	**WA → EN**		**EN → WA**	
Model \ Score	**chrF3**	**BLEU**	**chrF3**	**BLEU**
NLLB	47.8	20	34.9	2.2
+ EA	**50.1**	20.3	36.4	2.2
+ sWA-mono$_{\text{NLLB + EA}}$	49.8	**20.7**	45.6	7.8
+ sWA$_{\text{NLLB + EA}}$	49.8	20.5	**51.5**	**13.5**
+ WA	57.2	29.4	54	17
+ EA + WA	57.4	29.3	**54.2**	**17.1**
+ {WA, sWA-mono$_{\text{NLLB + WA}}$}	**57.7**	**29.8**	**54.2**	16.6

Table 3: Results on General Performance

The results shown in Table 3 are presented in two sections. The upper section contains the models without any genuine WA parallel data, i.e. the zero-resource case; whereas the lower section includes the models that are trained with genuine WA parallel data, i.e. the low-resource case.

In the zero-resource case, the results in each translation direction yield a different picture. When translating into English, the baseline's score is already relatively high, indicating that the system can somewhat handle WA input and capture a portion of its meaning correctly. This is also a confirming information to the WA-speaking community's intuition of using EA-trained MT models for translating WA texts. Additional EA finetuning results in very slight increases in both directions.Interestingly,

the increase in chrF3 is comparably larger than in BLEU score.

Training with a synthetic dataset generated by NLLB + EA does not bring much of an improvement in this direction, since the direction of the data generated is the same as the evaluated, meaning the generator system was already capable of generating those sentences. Feeding them into the system again will not bring much new information. Coming to the opposite translation direction, the baseline and EA-finetuned models perform very poorly because these have no knowledge of generating a WA sentence. Even though both languages have a substantial share of vocabulary, they use different orthographies (e.g. the word for "then/afterwards" is spelled in WA as "յետոյ" [hedo] whereas in EA as "հետո" [heto]), which means even if the correct word is chosen, the orthographical difference results into mismatches for both chrF3 and BLEU. Introducing WA through synthetic examples seems to increase the performance in this direction because the system sees genuine WA sentences even though there are mistakes in the mapping of meaning (e.g. the present tense indicator of WA corresponds to the future tense indicator of EA). The mistakes could also be reasoned with domain mismatches since the monolingual data is from a different domain than the parallel training and test data. Having the same domain in training data as the test data results in a doubling of BLEU scores and a nearly 6-point increase in chrF3. Meanwhile, the scores of this model come near to the scores of the models in the supervised case. This is an important insight, showing that using only monolingual data and pretrained EA models, one can generate synthetic training datasets and the models trained with it can reach comparable performance levels with the WA-trained models. This convergence additionally hints at the importance of matching domains in test and training data. Furthermore, the domain mismatch seems to be of more importance when translating into the low-resource language than when translating out of it.

In the supervised case, we see an average increase of 9 BLEU / 7 chrF3 points in WA → EN direction and a 4 BLEU / 3 chrF3 points increase in the opposite direction from the best model in the zero-resource case. In both directions, we do not see considerable improvements when additional data is introduced. As indicated in the zero-resource case the additional finetuning on EA did not change the model's knowledge much. This is again confirmed here, rendering the models NLLB + WA and the doubly finetuned NLLB + EA + WA the same. The increase seen with the introduction of synthetic examples in EN

Evaluated on: WA-Bible-test				
Direction	**WA → EN**		**EN → WA**	
Model \ Score	**chrF3**	**BLEU**	**chrF3**	**BLEU**
NLLB	50.2	23.7	34.4	2.6
+ WA-Bible	61	36.9	58.4	22
+ WA-Wiki	28	5.6	28.3	1
+ EA-Bible	40.5	12.9	32.6	1.6
+ EA-Wiki	39.9	14.4	26.5	0.4

Table 4: Results on the Effect of Domain vs. Language

5.3. Domain vs. Language

The common sense will suggest that the model that has been trained with the matching language and domain as the test set will get the highest and the one that has been trained with both mismatching domain and language will get the lowest result in both directions. The interesting part of the experiment is how the other models rank up; additionally, how the baseline model performs in this altogether, as well as the relative performances of the models against the baseline.

Surprisingly, as seen in Table 4 the intuition fails in one of the cases. In WA → EN direction, the lowest performance comes from NLLB + WA-Wiki, where the training data has matching language; whereas NLLB + EA-Wiki, the model that has been trained with wholly mismatching training data and was expected to come last, ranks second in BLEU scores. This is probably caused, because the EA-Wiki dataset contains information about the Bible, however in the opposite direction even if the knowledge is there it cannot be mapped onto the correct WA outputs. In both translation directions, performance drops below the baseline when a mismatch is present in the training data. The drop in performances has different severities in both directions. When translating into English, some portion of WA input is acknowledged correctly, which was already highlighted in the previous experiment. In the opposite direction, the performance drops severely. In the case of mismatching languages, the models never see any WA sentence and therefore have no information about generating one. In the case of NLLB + WA-Wiki, the drop is probably caused by the stylistic differences between the Bible and Wikipedia articles. As a general result, in both directions, the combination of matching domain-mismatching language has better results than matching language-mismatching domain, which tells us information gained from the matching domain has more importance than from the texts of the same language having a different domain. One can argue that this is only the case for the Bible subset. To confirm this, the WA-English parallel corpus must be extended with the datasets which have the same domains as the EA-English parallel

corpora.

6. Conclusion

In this work, we built the first NMT model and the parallel corpus of the endangered Western Armenian and English. We surveyed the WA's place in NLP research by listing the related work. We listed available resources of WA as well as some tips on how to extend the search on finding WA sources. We created the first WA-English parallel corpus with a total of approximately 147k examples covering a fair range of domains, whose copyright-free section of 52k examples was shared publicly. We investigated the WA translation performance in zero-resource and supervised settings. We found out that when translating into English, the EA-trained models could capture a considerable portion of WA input and map to the correct English outputs. In EN → WA direction, EA-trained models perform very poorly, since the models do not see any kind of WA sentence and therefore do not know how to generate it, however training on synthetic parallel data originating from monolingual WA data yields performance levels that are near to the supervised case. In the supervised case, additional data alongside genuine WA parallel data did not bring much of an improvement. In a separate experiment, we found out that information from the matching domain is generally more important than matching language. Any kind of mismatch in training data resulted in more severe performance drops when translating into WA than into English. The best model in translation achieves a BLEU score of 29.8 in WA → EN and 17.1 in EN → WA direction.

6.1. Future Work

The experiments have shown the significant effect of parallel training data. Therefore, the work on creating a parallel WA-English corpus is only a starting point. For high-quality WA translation, additional parallel resources should be investigated.

With this work, we aim to attract the interest of researchers for the endangered Western Armenian language and hope for more collaborative works. On that occasion, we want to highlight the need for additional tools for WA data collection. As mentioned previously, the quality of the OCR on WA texts was poor, an improvement here would result in more efficient processing of WA printed text and therefore a faster data collection process. Additionally, including WA word embeddings in multilingual embedding spaces would enable mining parallel data in many languages coupled with WA.

7. Acknowledgements

We would like to express our sincerest gratitude to the Calouste Gulbenkian Foundation for their support in this project.

8. Bibliographical References

2022. Մեծ Հայք համազգային ցանց. (National Center of Communication and Artifical Intelligence Technologies).

2024. Վիճակագրութիւն — Ուիգիփեդիա. (Statistics - WA Wikipedia).

R. Ananthakrishnan, Pushpak Bhattacharyya, M. Sasikumar, and Ritesh M. Shah. 2007. Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *Icon*, 64.

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, Marseille, France. European Language Resources Association.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Maryam Borjian. 2017. *Language and globalization: An autoethnographic approach*. Taylor & Francis.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

George L. Campbell. 2003. *Concise compendium of the world's languages*. Routledge.

Haroutioun Hovanes Chakmakjian. 1914. *Armeno-American Letter Writer, Containing a Large Variety of Model Letters Adapted to All Occasions: Letters of Friendship, Letters of Congratulation and Condolence, Letters of Love, Business Letters*. EA Yeran.

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, Marseille, France. European Language Resources Association.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation.

Council of Europe. 1992. States parties to the european charter for regional or minority languages and their regional or minority languages.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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, Marseille, France. European Language Resources Association.

Anaid Donabedian-Demopoulos. 2018. Middle east and beyond-western armenian at the crossroads: A sociolinguistic and typological sketch.

Ethnologue. 2023. Languages of the world.

Mikel L. Forcada, Gema Ginest\'\irez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

Inalco. 2021. Le projet prc "dalih - digitizing armenian linguistic heritage" est lauréat de l'aapg 2021 de l'anr.

Mike Izbicki. 2022. Aligning word vectors on low-resource languages with wiktionary. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 107–117, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8066–8073.

L. Khachatryan. 2011. Formalization of proper names in the western armenian press. In *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 Interna- tional Conference*, pages 75–85, Newcastle, UK. Cambridge Scholars Publishing.

L. Khachatryan. 2012. An armenian grammar for proper names. In *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2012 International Conference*, Newcastle, UK. Cambridge Scholars Publishing.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless?

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work?

Jeroen Ooms. 2023. tesseract: Open source ocr engine.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Georg Rehm and Andy Way. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Nature.

Nipun Sadvilkar and Mark Neumann. 2020. Pysbd: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning.

SIL International ISO 639-3 Registration Authority. 2017. Registration authority decision on change request no. 2017-023: to create the code element [hyw] for western armenian.

Max Silberztein. 2005. Nooj: a linguistic annotation system for corpus processing. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 10–11.

Aaron Smith, Christian Hardmeier, and Joerg Tiedemann. 2016. Climbing mont bleu: The strange world of reachable high-bleu translations. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 269–281.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pages 341–351.

Team-NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation.

United Nations Educational, Scientific and Cultural Organization. 2010. *Atlas of the World's Languages in Danger*, 3. ed., entirely rev., enlarged and updated. edition. Paris, France.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation.

Xiang Yu, Ngoc Thang Vu, and Jonas Kuhn. 2020. Ensemble self-training for low-resource languages: Grapheme-to-phoneme conversion and morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 70–78, Online. Association for Computational Linguistics.

9. Language Resource References

Nisan Boyacioglu and Hossep Dolatian. 2020. *Armenian Verbs: Paradigms and verb lists of Western Armenian conjugation classes*. Zenodo.

Dolatian, Hossep and Swanson, Daniel and Washington, Jonathan. 2022. *GitHub – apertium-hyw-corpus*.

Donabedian-Demopoulos, Anaid and Boyacioglu, Nisan. 2007. *La lemmatisation de l'arménien occidental avec NooJ*. Presses Universitaires de Franche Comté.

NooJ. 2023. *NooJ – Resources*.

Yavrumyan, Marat M. 2023. *UD Western Armenian ArmTDP*.