ManNER & ManPOS: Pioneering NLP for Endangered Manchu Language

Sangah Lee, Sungjoo Byun, Jean Seo, Minha Kang
Department of Linguistics, Seoul National University
{sanalee, byunsj, seemdog, alsgk1123}@snu.ac.kr

Abstract

We present pioneering research in the realm of Natural Language Processing (NLP) for the endangered Manchu language. Recognizing the critical importance of linguistic preservation, we experiment with three language models – BiLSTM-CRF, BERT, and mBERT – for Named Entity Recognition (NER) and Part-of-Speech (POS) tagging tasks. Given the limited digitized Manchu text available, we augment the data using GloVe embeddings for the pre-training of BERT-based models. Remarkably, all models demonstrated outstanding performance, achieving over 90% F1 score in both NER and POS tagging tasks. Our research not only marks the first application of NLP on Manchu and the inaugural use of BERT-based models for the language but also stands as the first endeavor to employ Manchu for NER and POS tagging. To foster further exploration and applications in the field, we make our fine-tuning dataset and models available to the public. Through this research, we aim to underscore the significance of NLP in the protection and revitalization of low-resource languages.

Keywords: Manchu language, NER, POS tagging, Low-Resource Language, Endangered Language, Data Augmentation, BiLSTM-CRF, BERT, mBERT

1. Introduction

Natural Language Processing (NLP) has experienced a meteoric rise over the past few years, with the focal point of this development predominantly centering around English and other Indo-European languages. The momentum of advancements in these language domains is accelerating at an unprecedented rate. Consequently, there is an increasing disparity between the attention and resources dedicated to dominant languages and the low-resource languages.

Manchu, an extremely low-resource language, typifies this dichotomy. In addition, while the linguistic study of the Manchu language is vibrant, NLP research in Manchu remains unexplored. Our study embarks on the creation of transformer-based language models trained on Manchu, demonstrating commendable performance, specifically in the NLP tasks of Named Entity Recognition (NER) and Part-of-Speech (POS) tagging. Table 1 shows snippets of examples from the task datasets.

The main challenge to training a Transformer model with Manchu is the limited volume of digitized Manchu text data. To circumvent this, we employ the data augmentation method from Seo et al. (2023). We then develop model architectures including BiLSTM-CRF, multilingual BERT (mBERT), and BERT. Notably, we trained three separate BERT models from scratch using three versions of augmented data.

For NER, while precision was highest with BiLSTM-CRF, recall and F1 score peaked with BERT trained with half-augmented data, where half of the words in each sentence in the dataset are replaced with their respective synonyms via the GloVe embeddings. For POS tagging, BERT trained with half-augmented data overshadowed in the precision, recall, and F1 score, although BERT trained with full-augmented data matched in the recall.

For further research, we anticipate more rigorous and diverse transformer-based models utilizing digitized Manchu data via OCR. This augments prospects of branching into other NLP tasks, including Transformer-based machine translation. Moreover, we aim to make this research applicable to other studies focusing on low-resource languages. The tasks performed in this study, such as POS tagging and NER, will be beneficial for facilitating analyses that require less labor-intensive approaches for the yet unannotated corpora in the Manchu language.

Our contributions are as follows:

- Construction of Manchu NLP Task Datasets
- Training and evaluation of Manchu language models
- Performing an in-depth study on the endangered Manchu language

2. Related Works

Low-resource language model

Numerous endeavors have been made to develop language models for low-resource languages. For instance, AraBERT (Antoun et al., 2020) and PhoBERT (Nguyen and Nguyen, 2020) are BERT variations specialized for Arabic and Vietnamese

LREC-COLING 2024, pages 11030–11039

text	**label**
Named Entity Recognition	
te geli lio be wen i gisun de latuna habi , now also Liu Bo Wen GEN word DAT approach PRS.PRF “Now you are also obsessed with what Liu Bowen has said.”	O O BSC-B BSC-I BSC-I O O O O O O
hesihe de buce he seme , Hesihe DAT die P.PST COMP , “that he has died in Hesihe region,”	PLC-B O O O O O
POS tagging (simplified tags)	
han ji fi geli afa me gai mbi kai , Khan come CVB.ANT again attack CVB.SIM take NPST PTCL , “Khan will truly come again and attack you to take (the castle).”	nv verb e nv verb e verb e ptcl mark
emu ciyandzung de duin niyalma , one Qianba DAT four person , “He (gave) a Qianba (proper name of a position) 4 persons,”	nv nv cm nv nv mark
POS tagging (original tags)	
wang_tsanjiyang be gai fi , Wang_Canjiang ACC take CVB.ANT , “Accompanying Wang Canjiang (proper name of a position)”	nn::prpn cm::acc verb::vv e::cve mark::comma
suwen be bibu mbi o , 2PL ACC let.exist NPST Q , “I will not let you exist (if you are still there.)”	pn::2pl cm::acc verb::vv e::fve.npst ptcl mark::comma

Table 1: Task data examples

respectively. Ogueji et al. (2021) introduced AfriBERTa, a Transformer-based multilingual model trained on low-resource African languages. This model has showcased promising performance in tasks like text classification and NER. Furthermore, Azunre et al. (2021) unveiled a BERT model exclusively trained for Akuapem Twi, aiming to enhance machine translation capabilities.

Models like BLOOM (Scao et al., 2022), which was trained on 46 distinct natural languages, multilingualBERT (Libovický et al., 2019) that generates sentence representations for 104 languages, and XLM-RoBERTa, a transformer-based language model trained on a hundred languages, represent multilingual language models that include a broad spectrum of language families.

Language Tasks in the Low-Resource Scenario

Monolingual natural language processing (NLP) task datasets have been developed for languages with limited linguistic resources: a paraphrased text set for Bangla (Akil et al., 2022), a dataset for sentiment analysis of major Nigerian languages (Muhammad et al., 2022), Bengali news dataset (Akash et al., 2023), and so on. Benchmarks that collect several tasks written in a specific low-resource language are like the following: Turkish[1], Lao[2], Hausa[3], and so on.

For the NER task, many studies have explored the utilization of pre-trained multilingual models or transfer learning techniques across languages, particularly in low-resource settings. Chen et al. (2021) fine-tuned XLM-RoBERTa with augmented datasets for Uyghur and Hungarian, while Sabane et al. (2023) investigated several monolingual and multilingual models for Hindi and Marathi. Dang et al. (2023) incorporated BERT and a lexical adapter for addressing challenges in Chinese. Furthermore, Torge et al. (2023) employed RoBERTa and high-resource languages from the same language family to tackle NER tasks in Upper Sorbian and Kashubian.

For the POS tagging task, numerous studies have employed neural network models including RNNs, LSTMs, and sequence-to-sequence systems, addressing languages such as Latin (Celano, 2020; Wróbel and Nowak, 2022), Romanian (Lőrincz et al., 2019), and Amharic (Hirpassa and Lehal, 2023). Recently, there has been a surge in research focusing on Transformer-based model architectures. Notably, research endeavors include POS tagging of Gurmukhi Punjabi utilizing the IndicBERT-BiLSTM architecture (Kumar and Sikka, 2023) and POS tagging of Korean leveraging a two-layer Transformer encoder alongside a novel syllable-based approach (Shin et al., 2023).

3. Data Construction

3.1. ManNER & ManPOS: Construction of Manchu NLP Task Datasets

The Manchu language is one of the Manchu-Tungusic languages used in Northern Asia. It was the major language in the Qing Dynasty, but now it is nearly extinct (Kim et al., 2008) and is primarily used in written form. However, there are several old literature written in the Manchu language. We

[1] https://github.com/GGLAB-KU/turkish-plu

[2] https://github.com/wannaphong/Awesome-Lao-NLP

[3] https://github.com/hausanlp/hausanlp

will use the term ‘Manchu language’ to refer to this written form of Manchu in our study.

The main source of our NER task and POS tagging is the morphologically annotated corpus of Manwen Laodang Taidzu, introduced by Choi et al. (2023)[4]. Manwen Laodang is a piece of literature containing the history of the uprising of the Jurchen and Manchu people. The literature is written in the Manchu language using the Manchu script. Choi et al. (2023) transcribed this book using Möllendorff (1892)’s romanizing system and annotated each morpheme.

This corpus has the format of JSON, and the total number of morphemes in the corpus is 252,645. The names and structures of the tags in the corpus are provided from Choi et al. (2023), and the full tag list can be found in Table 11. They are classified as content words, functional words, and punctuation marks. Each classification includes some classes, and they have several categories, which are denoted by a double semicolon (‘::’). The example of the tagging for the Manchu sentence ‘*tumen cooha be unggifi tosoho,*’ (“He sent 10,000 soldiers to block the road.”) is shown in Table 2. The meanings of the notations in Table 2 can be found in Table 3.

form	analyzed & annotated
tumen	tumen/nv::nv
cooha	cooha/nv::nv
be	be/cm::acc
unggifi	unggi/verb::vv + fi/e::cve
tosoho	toso/verb::vv + ho/e::pst.ptcp
,	,/mark::comma

Table 2: Examples of annotation for Manchu sentences in Choi et al. (2023)’s corpus

tag	class	category
nv::nv	non-verbal	non-verbal
cm::acc	case marker	accusative
verb::vv	verb	stem
e::cve	ending	converb
e::pst.ptcp	ending	past participle
mark::comma	marker	comma

Table 3: Meanings of the tags in the example of Table 2

Proper nouns, the focus of our NER task, are annotated with the following tags: “prpn” for basic proper nouns; “prpn.person” for personal proper nouns; “prpn.plc” for place proper nouns, as shown in Table 4. The numbers of each kind of pronouns are as follows: 5,622 basic proper nouns; 4,417 personal proper nouns; 1,145 place proper nouns. The sum of these is 11,184 (Choi et al., 2023). In NER task, we change “prpn” to “BSC” (BaSiC proper noun), “prpn.person” to “PER” (PERsonal proper noun), and “prpn.plc” to “PLC” (PLaCe proper noun). The labeling of Named Entity Recognition (NER) tags adheres to the widely used BIO system, as commonly reported in the literature.

The NER and POS tagging task datasets are provided via the GitHub repository.[5]

Manchu	Romanized form	tag
ᠪᠠᠪᠠᡳ ᡝᡶᡠ	babai efu	BSC
ᠪᡠᠵᠠᠨᡨᠠᡳ	bujantai	PER
ᡥᠣᡳᡶᠠ	hoifa	PLC

Table 4: Examples of proper nouns in Manchu

3.2. Data Augmentation for a Monolingual Manchu Corpus

We collect digitized, romanized monolingual Manchu data for the purpose of training BERT models. Due to the limitation of digitized Manchu text data, we employ the data augmentation method from Seo et al. (2023). This approach involves training GloVe embedding models with two versions of the dataset: one with sentences of at least 3 words and another with sentences of at least 5 words. We use various window sizes during training (1, 3, 5, 7, and 10), resulting in 10 distinct GloVe embeddings.

For each word in the dataset, we find the most similar word predicted by each GloVe model. The word with the highest frequency among the 10 models is chosen as the synonym. We then replace a word in each sentence from text data with this identified synonym. This process generates two augmented dataset versions. The first replaces as many words as possible with synonyms(‘full augmentation’), significantly increasing the dataset size relative to the average sentence length. The second version replaces half of the words in each sentence with their respective synonyms(‘half augmentation’), resulting in a dataset about half the size of the first method.

For fairness, we shuffle the sentences in the original data and then split it into test data and data for augmentation. The split data is then augmented into two separate versions, half augmentation and full augmentation respectively. Additional details of dataset size can be found in Table 5.

[4] The corpus can be downloaded through this GitHub repository: https://github.com/Kkamakpyel/manwenlaodang

[5] https://github.com/sanajlee/Manchu-NLP

augmentation	dataset size (# of sentences)
no augmentation	195,611
half augmentation	2,698,159
full augmentation	5,207,069
test data	19,541

Table 5: The size of each dataset

4. Training Manchu LMs

4.1. Monolingual Manchu BERT

In an inaugural endeavor to utilize the Transformer architecture on Manchu language, we embark on training and comparatively assessing three versions of BERT. These three models exhibit identical architectural configurations and training objectives, differing solely in the volume of training data. To elaborate, the three versions of training data encompass: the entirety of accessible monolingual Manchu data without any augmentation; data with half augmentation; and data with full augmentation. These models, which employ BERT-base configuration, are trained for 10 epochs with the following hyperparameters: vocabulary size of 25,000 using WordPieceTokenizer, maximum length of 512, and training batch size of 10.

4.2. Adaptation of Multilingual BERT to Manchu corpus

We attempt to adapt the well-trained multilingual BERT model to the unseen Manchu language. Such a multilingual model is expected to have learned general linguistic knowledge from various languages in its training corpus. Then we expect the model to additionally learn the properties of Manchu and align them to its existing embedding space. We employ the checkpoint of mBERT (bert-multilingual-uncased)[6] and continually pre-train it with our Manchu corpus.

We utilize the 'full augmentation' version of monolingual Manchu texts as the corpus for continually training mBERT. Additionally, tokens to expand the mBERT vocabulary are selected by training a WordPiece tokenizer on the corpus. We get 10,000 tokens from the tokenizer and add 8,325 tokens which are not included in the original mBERT tokenizer. The expanded tokenizer includes 114,204 tokens in its vocabulary.

We train the model for one epoch, with all the configurations and settings following those of the original mBERT. For further training, the hyperparameters include the sequence max length of 512 and training batch size of 16.

4.3. BiLSTM-CRF Models for Manchu NLP Tasks

In our study, we employ the Bi-LSTM-CRF model, which integrates a bidirectional Long Short-Term Memory (Bi-LSTM) network with a Conditional Random Field (CRF) layer (Lafferty et al., 2001). By employing the strengths of both Bi-LSTM and CRF, this combined architecture not only leverages contextual features from the entire input sequence but also delivers improved tagging accuracy. It is noteworthy that the BiLSTM-CRF configuration has gained widespread adoption in tasks such as POS tagging, chunking, and various NER datasets.

In BiLSTM-CRF, the embedding layer transforms word indices into continuous vectors. We set the embedding dimension e to 256. The size of the hidden state, represented by h, is also configured to 256. To reduce overfitting, we introduce a 0.5 dropout rate. We employ the 'full augmentation' version of monolingual Manchu texts as the training dataset for this model over 10 epochs, utilizing a training batch size of 32. The maximum sequence length is adjusted to accommodate the longest sequence within the training dataset.

5. Experiments

5.1. Experiment Settings

For task inference, we split each dataset into training, validation, and test sets with a ratio of 8:1:1. As a result, the constructed datasets include 27,510, 3,519, and 3,531 examples in the training, validation, and test sets, respectively, for all the tasks we provide in this paper.

We perform each task experiment with six different models: BiLSTM (baseline), BiLSTM-CRF, the mBERT-based model that we continually train with the monolingual Manchu texts (mBERT), and the three versions of pre-trained monolingual Manchu language-based BERT models with the different portions of word replacement for data augmentation (no aug, half aug, full aug).

As a baseline, we opt for the basic bidirectional Long Short-Term Memory (BiLSTM) model. This model is trained with configurations identical to those outlined for our BiLSTM-CRF model in Section 4.3, with the exception of the maximum sequence length set to 50.

Especially, we fine-tune the monolingual BERT models and the Manchu-adapted mBERT model that we mentioned in sections 4.1 and 4.2, to perform the NER and POS tagging tasks that we provide. We employ and modify the basic structure of the fine-tuning codes from KoELECTRA (Park, 2020), with hyperparameters such as the max sequence length of 128, training batch size of 32, and learning rate of 5e-5. The training procedure for each task is repeated for 20 epochs.

[6] https://github.com/google-research/bert/blob/master/multilingual.md

5.2. Named Entity Recognition

Model	Precision	Recall	F1
BiLSTM	86.23	72.02	77.90
BiLSTM-CRF	**93.53**	93.91	93.72
mBERT	92.47	91.64	92.05
no aug	91.41	92.45	91.92
half aug	93.23	**94.97**	**94.09**
full aug	92.59	94.32	93.45

Table 6: Overall Performance for NER

As shown in Table 6, the half-augmented monolingual BERT model ('half aug') shows the best F1 score for Named Entity Recognition, although the BiLSTM-CRF model obtained the best precision score. Then the fully augmented BERT model ('full aug') shows slightly lower performance than those of the best models. Among the pre-trained monolingual BERT models, the models that were trained with augmented datasets show better performance than the model trained on the original Manchu texts without any data augmentation process. All these outcomes surpass the baseline results obtained with the simple BiLSTM model, thereby demonstrating the discriminative capacity of our dataset, which underscores the complexity of the task.

NER Label	Precision	Recall	F1
BiLSTM			
BSC	88.05	66.67	75.88
PER	78.55	88.34	83.16
PLC	93.84	69.47	79.84
Overall	86.23	72.02	77.90
BiLSTM-CRF			
BSC	93.39	93.24	93.31
PER	91.58	95.07	93.29
PLC	100.00	93.33	96.55
Overall	93.53	93.91	93.72
mBERT			
BSC	94.74	92.75	93.73
PER	87.94	89.91	88.91
PLC	96.82	92.12	94.41
Overall	92.47	91.64	92.05
BERT (half aug)			
BSC	95.04	95.65	95.35
PER	90.32	94.17	92.21
PLC	94.55	94.55	94.55
Overall	93.23	94.97	94.09

Table 7: Performance on each tag class for NER

In the NER task, as in Table 7, the prediction accuracy for proper nouns is nearly 0.95, but that of personal proper nouns is relatively low. The errors typically manifest in various ways, such as PER-B being mispredicted as O, O being mispredicted as PER-B, and BSC-I and BSC-O also being mispredicted as O. These tendencies were observed consistently across all models and settings in our study.

This is due to the characteristics of personal proper nouns and the fundamental issues in the annotated corpus. Many personal names in the Manchu language are derived from bare nouns. For example, *hashū*, a nickname for the Ula tribe, originally means "leaf side."Additionally, multi-word proper nouns in the corpus for our study were not split into separate forms. It can be easily inferred that common nouns like defining official positions are also classified as BSC.

In the baseline scenario, the simple BiLSTM model exhibits notably low recall scores for the classes BSC and PLC, and also demonstrates lower performance on the PER label compared to the other proposed models.

5.3. Part-of-Speech Tagging

Simplified Tags

Model	Precision	Recall	F1
BiLSTM	81.15	76.20	78.26
BiLSTM-CRF	99.48	97.39	98.27
mBERT	99.34	99.33	99.33
no aug	99.78	99.78	99.78
half aug	99.49	99.48	99.48
full aug	**99.82**	**99.80**	**99.81**

Table 8: Overall Performance for POS Tagging (simplified tags)

Table 8 reports that nearly all of the models exhibit perfect performance in the POS tagging tasks for all classes. However, as in Table 9, the only exception is the prediction of the 'sfx' tag, with a recall score of 0.75. This may be due to the homogeneous form of the morpheme tagged with 'sfx' and the morpheme tagged with other labels. For example, *bu*, tagged with 'sfx,' serves as the causative-passive suffix. Yet, *bu* is also the stem of a common verb meaning 'to give.' However, the frequency of mismatched 'sfx' tags is only four.

In the case of the baseline BiLSTM model, it correctly predicts none of the instances for the tag 'sfx.' Nevertheless, it demonstrates notable performance across other tags, akin to the performance exhibited by other proposed models.

The high tagging performance for simplified tags can be attributed to several factors. Firstly, the number of simplified tags is smaller than that of the full tags, and they do not require consideration of additional information such as negation or noun quality (e.g., proper noun, etc.) because these details are already incorporated within the simplified tags.

Model	BiLSTM			BiLSTM-CRF			BERT			mBERT		
POS Tag Label	**Precs.**	**Recall**	**F1**	**Precs.**	**Recall**	**F1**	**Precs.**	**Recall**	**F1**	**Precs.**	**Recall**	**F1**
cm	99.63	99.97	99.80	99.89	99.94	99.91	99.97	99.97	99.97	99.92	99.92	99.92
comp	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
e	99.73	99.79	99.76	99.95	99.93	99.94	99.88	99.83	99.86	99.88	99.88	99.88
mark	100.00	100.00	100.00	99.92	99.79	99.85	100.00	100.00	100.00	100.00	100.00	100.00
nn	96.36	95.51	95.94	96.65	98.83	97.73	96.93	98.31	97.62	95.61	95.61	95.61
nv	99.26	99.79	99.53	99.79	99.44	99.61	98.92	98.69	98.80	98.48	98.45	98.47
pn	99.67	79.52	88.46	98.11	99.18	98.64	99.72	100.00	99.86	100.00	99.72	99.86
ptcl	98.99	65.63	78.93	100.00	99.04	99.52	100.00	99.06	99.53	97.25	100.00	98.60
sfx	0.00	0.00	0.00	100.00	75.00	85.71	100.00	75.00	85.71	100.00	75.00	85.71
verb	99.02	98.08	98.55	99.90	99.88	99.89	99.88	99.79	99.84	99.79	99.77	99.78
Overall	81.15	76.20	78.26	99.48	97.39	98.27	99.49	99.48	99.48	99.34	99.33	99.33

Table 9: Performance on each tag class for POS Tagging (simplified tags)

Model	Precision	Recall	F1
BiLSTM	61.48	61.41	60.44
BiLSTM-CRF	92.09	90.96	91.45
mBERT	**98.85**	**98.86**	**98.86**
no aug	98.65	98.61	98.63
half aug	98.84	98.81	98.82
full aug	98.83	98.83	98.83

Table 10: Overall Performance for POS Tagging (original tags)

Secondly, in the Manchu language, when a sentence is correctly tokenized, the surface forms and parts of speech almost exhibit a one-to-one correspondence. While this assertion may not hold true for the full tags, the simplified tags notably exhibit this characteristic, attributed to the nature of the Manchu language, which seldom incorporates irregular forms of inflection. As a result, our model can easily carry out the simplified tagging task.

Original Tags

As illustrated in Table 10, all models evaluated for the POS tagging task using the complex original tagset achieve excellent performance, with F1 scores exceeding 90, except for the baseline BiLSTM model.

While the BiLSTM-CRF model demonstrates comparatively lower performance compared to other models, all the BERT-based models show precision, recall, and F1 scores over 98, and the models report similar score values to each other. Here, the mBERT-based model further adapted to the Manchu language obtain the best performance, being slightly higher than the other BERT-based models.

Among the monolingual BERT models, the models trained on the augmented corpus show slightly better performance than the model trained only on the original corpus. However, the difference in scores between the half-augmented and the full-augmented models does not seem to be significant.

Considering Table 11, it can be observed that all the models exhibit low performance in tagging the following classes: ‘e::fve.imperative’ (imperative ending), ‘e::fve.prv’ (preventive ending), ‘e::fve.prs.prf.neg’ (negation of present perfect), ‘e::npst.ptcp.neg’ (negation of non-past participle), ‘e::pst.ptcp.neg’ (negation of past participle), and ‘nn::prpn.person’ (personal proper noun). mBERT-adapt model shows perfect score for ‘e::fve.imperative.’ BiLSTM-CRF shows a low score for ‘mark::rparen’ (right parentheses), despite exhibiting 100% performance for all other classes except ‘nn::prpn.person.’

The mismatch of ‘e::fve.imperative’ in BERT can likely be attributed to certain irregular forms of the imperative in the Manchu language. In particular, verb stems containing the substring *ha*, which is one of the allomorphs of the past participle ending, might play a crucial role in these errors. For example, *daha-*, meaning ‘to surrender,’ contains *ha* in its stem. Additionally, *baha-*, meaning ‘to get,’ shares the same form for both imperative and past participle. This similarity could lead to incorrect parsing in the original corpus.

The errors in tagging ‘e::fve.npst.ptcp.neg’ and ‘e::fve.pst.ptcp.neg’ may be attributed to words that contain the substrings *hakū* and *rakū*, which originally correspond to each of these tags. For example, *undurakū* and *cihakū* are predicted with these tags, but the correct tags should be ‘nv::nv’ and ‘nv::adj.’ The remaining issue, the mismatch of ‘nn::prpn.person,’ can be comprehended in the same context as the explanation of errors in the NER task (as seen in section 5.2). As for the mismatch of ‘e::fve.prv,’ ‘e::fve.prs.prf.neg’ and ‘mark::rparen,’ it primarily arises from their significantly low frequency, which are only 117, 44 and 26 times. Furthermore, ‘mark::unknown’ appears merely three times in the training set and is absent in the validation and test sets.

5.4. Analysis

In Sections 5.2 and 5.3, we present the performance metrics of the BiLSTM-CRF and BERT-based models, detailing the Precision, Recall, and F1 scores for each label. In this section, we evaluate and discuss the classification capabilities of these models, based on their inference results.

Model	BiLSTM			BiLSTM-CRF			BERT			mBERT		
POS Tag Label	Precs.	Recall	F1	Presc.	Recall	F1	Precs.	Recall	F1	Precs.	Recall	F1
cm::												
abl	66.98	93.85	78.17	100.00	98.35	99.17	99.20	100.00	99.60	99.17	96.77	97.96
acc	99.57	100.00	99.78	99.89	100.00	99.95	99.90	100.00	99.95	99.90	100.00	99.95
dat	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.80	100.00	99.90
gen	99.74	100.00	99.87	100.00	100.00	100.00	100.00	100.00	100.00	99.80	100.00	100.00
comp	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
e::cve	99.97	85.74	92.31	99.88	99.94	99.91	99.94	99.94	99.94	99.76	100.00	99.88
e::fve.												
imperative	100.00	100.00	100.00	100.00	100.00	100.00	86.50	85.64	86.07	100.00	100.00	100.00
npst	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	99.60	100.00	99.80
opt	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
prs.prf	99.46	96.83	98.12	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
prs.prf.neg	0.00	0.00	0.00	100.00	100.00	100.00	50.00	33.33	40.00	50.00	33.33	40.00
prv	0.00	0.00	0.00	100.00	100.00	100.00	50.00	33.33	40.00	100.00	92.86	96.30
pst.ipfv	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
e::npst.												
ptcp	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
ptcp.neg	99.56	100.00	99.78	100.00	100.00	100.00	81.30	85.39	83.30	87.84	89.04	88.44
e::pst.												
ptcp	99.97	99.92	99.94	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
ptcp.neg	100.00	97.80	98.89	100.00	100.00	100.00	53.70	47.54	50.43	61.11	54.10	57.39
mark::												
circle	100.00	100.00	100.00	100.00	100.00	100.00	95.69	97.56	96.62	100.00	100.00	100.00
comma	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
lparen	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
period	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
rparen	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00
unknown	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nmlz::nmlz	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
nn::												
prpn	38.64	95.09	54.95	97.05	87.54	92.05	99.67	100.00	99.83	99.83	99.50	99.67
prpn.person	97.26	85.73	91.14	86.96	96.77	91.60	89.41	95.05	92.14	89.20	93.02	91.07
prpn.plc	96.01	79.29	86.86	95.12	97.50	96.30	96.25	93.33	94.77	94.51	93.94	94.22
nv::adj	98.93	94.86	96.85	98.13	98.59	98.36	97.55	93.62	95.55	93.58	92.95	93.27
nv::nv	99.28	99.77	99.53	99.48	99.63	99.56	98.44	98.03	98.24	98.14	97.89	98.02
pn::												
1pl.excl	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1pl.incl	32.49	84.91	47.00	100.00	82.35	90.32	100.00	82.35	90.32	100.00	82.35	90.32
1sg	0.00	0.00	0.00	99.29	100.00	99.64	100.00	99.30	99.65	100.00	99.30	99.65
2pl	100.00	94.98	97.43	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
2sg	98.81	40.10	57.04	100.00	98.55	99.27	100.00	100.00	100.00	100.00	100.00	100.00
3pl	0.00	0.00	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
3sg	97.14	95.77	96.45	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
ptcl	97.76	77.90	86.71	100.00	99.04	99.52	99.06	99.06	99.06	98.15	100.00	99.07
sfx	0.00	0.00	0.00	100.00	75.00	85.71	100.00	75.00	85.71	100.00	75.00	85.71
verb::												
vv	99.34	95.53	97.40	99.93	99.95	99.94	99.88	99.83	99.86	99.81	99.81	99.81
vv.npst	0.00	0.00	0.00	100.00	96.30	98.11	100.00	100.00	100.00	96.55	100.00	98.25
Overall	61.48	61.41	60.44	92.09	90.96	91.45	98.84	98.81	98.82	98.85	98.86	98.86

Table 11: Performance on each tag class for POS Tagging (original tags)

Model Performance

The BiLSTM-CRF model shows comparable performance with other models in the NER and POS tagging (with simplified tags) tasks while reporting performance drop in the POS tagging with original tags. We guess that the original POS tags are much more diverse and complicated than the simplified version, letting the model perform poorly. Especially, the BiLSTM-CRF model has many of its errors with specific tags: 'nn::prpn' and 'nn::prpn.person.' These two tags are subcategories of the simplified tag 'nn,' where the model shows a comparable F1 score of 97.73. The model may have difficulties with distinguishing the detailed subcategories of noun morphemes ('nn'). The model also shows the F1 score of 0.00 for the tag 'mark::rparen,' but the test set included only three morphemes annotated with the tag.

When comparing the three versions of monolingual BERT models, the models trained on the augmented dataset report higher performance in all three tasks. However, the performance difference between half-aug and full-aug models is not very significant. The training datasets of both models are different only in the ratio of words replaced by the GloVe-based similar words during the data augmentation process. Here, when increasing the ratio of word replacement, more copies of similar sentences with replaced words of the original dataset will be created, while retaining the other words and overall syntax and semantics of the sentence. This may degrade the diversity of data examples in the dataset, harming the ability of the language model to learn linguistic knowledge of Manchu texts. According to our experiment results, replacing words over half of the sentence length does not affect the downstream performance so much, even though the full augmentation setting constructs the largest scale of a corpus that we can utilize. Thus, we plan to compose a monolingual dataset of more diverse linguistic phenomena and contexts, utilizing different methods for data augmentation and external tools, to improve Manchu language modeling.

The mBERT-based model further adapted to Manchu texts reports the best performance in the

POS tagging task with original detailed tags, while showing a slight difference with the monolingual BERT-based models. In the POS tagging task with simplified tags, the model also shows comparable results with the other models. However, it reports the worst performance in the NER task, even worse than the non-Transformer BiLSTM-CRF model. The performance drop from those of other models is the largest (about 1%p) except for the POS tagging (original) performance of the BiLSTM-CRF model.

Properties of Manchu Grammar

The mistakes made by these models can be comprehended by considering the following characteristics of Manchu morphemes. First, in the Manchu language, nouns and adjectives share the same form and are not distinguished in their usage. Such adjectives can be referred to as 'nonverbal adjectives.' In languages that feature nonverbal adjectives, these adjectives are encoded with nonverbal elements (Stassen, 2013). The error of confusing nouns and adjectives may stem from this characteristic of Manchu adjectives.

Secondly, in the Manchu language, certain case markers or suffixes share a uniform form with other morphemes. For instance, *be* is primarily used as the accusative case marker, but it can also serve as the first-person plural exclusive pronoun. Similarly, *ci* functions as the ablative case marker, but it is also used as the conditional converb ending in Manchu (Gorelova, 2002). Consequently, these homogeneous forms can lead to mismatches described before. The error type caused by *bu* reported in section 5.3 can be understood in a similar manner.

Finally, there are some morphemes that have been incorrectly tagged in the original corpus. For instance, *aihū*, which means 'female sable,' is tagged as 'nv::nv' (non-verbal) when it should be tagged as 'nv::prpn.person' (personal proper noun). *demtu*, *kekuhe*, and *yungge*, which are proper nouns in this text, are tagged as 'nv::nv.' In our model, we tag these nouns as 'nv::prpn.person.' Therefore, our model assigns more accurate predicted tags compared to the original corpus. This issue reported so far could potentially be addressed in a future study, which will explore improvements to the Manchu language tokenizer.

Analysis of Continual Training of mBERT

Table 12 shows the performance of mBERT models comparing the effect of adaptation to the Manchu texts. In the table, 'original' depicts the vanilla mBERT model, while 'adapt' represents our continually-trained mBERT-based model. Surprisingly, the vanilla mBERT does not report especially lower performance than the adapted model.

Task	Model	Precision	Recall	F1
NER	original	**92.48**	**92.78**	**92.63**
	adapt	92.47	91.64	92.05
POS(sim)	original	**99.43**	**99.46**	**99.45**
	adapt	99.34	99.33	99.33
POS(org)	original	98.84	98.82	98.83
	adapt	**98.85**	**98.86**	**98.86**

Table 12: Performance comparison of the vanilla mBERT and continually trained mBERT models

For all three tasks, we observe comparable results with the two models. Exceptionally, the vanilla mBERT model shows a very low F1 score of 16.67 for the detailed original POS tag e::fve.prv.

We guess that further training of mBERT may cause catastrophic forgetting during adaptation to an unseen language. The general linguistic ability of the model may be harmed, degrading the downstream performance of the model. In addition, it may not be enough to adapt a language model trained on a large-scale corpus with a large vocabulary including over 100K tokens by further training it on a relatively small dataset. We may attempt to train and adapt the model using a larger Manchu corpus for longer steps.

6. Conclusion

In this paper, we presented the first NLP task datasets for the low-resource endangered Manchu language. We constructed datasets for Named Entity Recognition and Part-of-Speech tagging based on the morphologically annotated Manchu corpus, Manwen Laodang Taidzu, introduced by Choi et al. (2023). We trained the following language models as the task baselines: the task-specific BiLSTM-CRF models, the multilingual BERT model adapted to Manchu texts, and three versions of monolingual Manchu BERT models. The BERT-based models are trained on our monolingual Manchu corpora, which we augmented using GloVe embedding models due to the small size of the original corpus.

For future work, we plan to collect more Manchu texts for language modeling and several downstream tasks. Since most of the historical literature written in the Manchu language is in the form of scanned images, we could utilize techniques including OCR[7] to construct textual datasets sufficiently. Additionally, we would adapt and extend different types of Transformer-based language models such as XLM-RoBERTa and BLOOM for further performance improvement and generalization.

[7] https://github.com/tyotakuki/ManchuOCR

Bibliographical References

Abu Ubaida Akash, Mir Tafseer Nayeem, Faisal Tareque Shohan, and Tanvir Islam. 2023. Shironaam: Bengali news headline generation using auxiliary information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 52–67. Association for Computational Linguistics.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15. European Language Resource Association.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. English-twi parallel corpus for machine translation. *arXiv preprint arXiv:2103.15625*.

Giuseppe G. A. Celano. 2020. A gradient boosting-Seq2Seq system for Latin POS tagging and lemmatization. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–123, Marseille, France. European Language Resources Association (ELRA).

Siqi Chen, Yijie Pei, Zunwang Ke, and Wushour Silamu. 2021. Low-resource named entity recognition via the pre-training model. *Symmetry*, 13(5).

Woonho Choi, Sunghoon Jung, and Jeongup Do. 2023. Construction of the Manchu corpus: focusing on *Manwen laodang Taidzu*. *Altai Hakpo*, 33:67–87.

Xiaochao Dang, Li Wang, Xiaohui Dong, Fenfang Li, and Han Deng. 2023. Improving low-resource chinese named entity recognition using bidirectional encoder representation from transformers and lexicon adapter. *Applied Sciences*, 13(19).

Liliya M Gorelova. 2002. *Manchu grammar*. Brill Academic Publishers, Boston-Köln.

Sintayehu Hirpassa and G.S. Lehal. 2023. Improving part-of-speech tagging in amharic language using deep neural network. *Heliyon*, 9(7):e17175.

Juwon Kim, Jaeil Kwon, Dongho Ko, Yoonshin Kim, and Soonhwan Jeon. 2008. *Documentation of endangered Altaic languages*. Taehaksa, Seoul.

Dr Arun Khosla Kumar, Yogender and Dr Geeta Sikka. 2023. Gurmukhi punjabi part of speech tagging using indicbert-bilstm architecture. *International Journal of Emerging Technologies and Innovative Research*, 10.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Beáta Lőrincz, Maria Nu□u, and Adriana Stan. 2019. Romanian Part of Speech Tagging using LSTM Networks. In *Proceedings of the IEEE 15th International Conference on Intelligent Computer Communication and Processing*, Cluj-Napoca, Romania.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*.

Paul Georg von Möllendorff. 1892. *A Manchu Grammar, with Analyzed Texts*. American Presbyterian Mission Press, Shanghai.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126. Association for Computational Linguistics.

Jangwon Park. 2020. Koelectra: Pretrained electra model for korean. https://github.com/monologg/KoELECTRA.

Maithili Sabane, Aparna Ranade, Onkar Litake, Parth Patil, Raviraj Joshi, and Dipali Kadam. 2023. Enhancing low resource ner using assisting language and transfer learning. In *Proceedings of the 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1666–1671.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Jean Seo, Sungjoo Byun, Minha Kang, and Sangah Lee. 2023. Mergen: The first Manchu-Korean machine translation model trained on augmented data. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 118–124, Singapore. Association for Computational Linguistics.

Hyeong Jin Shin, Jeongyeon Park, and Jae Sung Lee. 2023. Syllable-based multi-posmorph annotation for korean morphological analysis and part-of-speech tagging. *Applied Sciences*, 13(5).

Leon Stassen. 2013. Predicative adjectives (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Sunna Torge, Andrei Politov, Christoph Lehmann, Bochra Saffar, and Ziyan Tao. 2023. Named entity recognition for low-resource languages - profiting from language families. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.