ManNER & ManPOS: Pioneering NLP for Endangered Manchu Language

Sangah Lee, Sungjoo Byun, Jean Seo, Minha Kang

{sanalee, byunsj, seemdog, alsgk1123}@snu.ac.kr

Department of Linguistics
Seoul National University

Manchu Language (*Manju Gisun*)

Manchu-Tungusic language family, South branch

- Used in Later Jin and Qing Dynasty as the common language
- native speakers: almost 0 (endangered language; Kim et al. 2008)

Target of the study: **Written Manchu**, not Colloquial Manchu

- Written and Colloquial Manchu is somewhat different language!
- There are several Manchu language literature written in Manchu script:
  - ✓ Manwen Laodang (Man. *Tongki fuka sindaha hergen-i dangse*)
  - ✓ The Romance of Three Kingdoms in Manchu (Man. *Ilan gurun-i bithe*)
  - ✓ The Plum in the Golden Vase (Man. *Jin Ping Mei bithe*)
  - ✓ etc.

ManNER & ManPOS

Based on Manwen Laodang Taidzu (Choi et al., 2023)


```
"id": "MMNTMADOC0001.b01.1",
"form": "tongki fuka sindaha hergen i dangse .",
"word": [
    {
        "id": "1",
        "form": "tongki",
        "ma": "tongki/nv::nv"
    },
    {
        "id": "2",
        "form": "fuka",
        "ma": "fuka/nv::nv"
    },
    {
        "id": "3",
        "form": "sindaha",
        "ma": "sinda/verb::vv + ha/e::pst.ptcp"
    },
    {
        "id": "4",
        "form": "hergen",
        "ma": "hergen/nv::nv"
    },
```

ManNER & ManPOS

Based on Manwen Laodang Taidzu (Choi et al., 2023)

Set a list of classes and categories

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

We utilize these classes and categories as the tags for NER and POS tasks.

e.g.) prpn (proper noun) → named entity tags

class → simplified POS tags

class + category → “original” POS tags

ManNER & ManPOS

Constructed datasets for NER and POS tagging

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

Can be downloaded through:

ManPOS: Part-of-Speech Tagging

Utilizing the analyzed tags from the reference set

Two versions of POS tagging dataset: Simplified and Original

- Simplified: using only the upper “class” information
- Original: using both of the “class” and “category” information

Analyzed Morpheme	Simplified Tag	Original Tag
tumen	nv	nv::nv
cooha	nv	nv::nv
be	cm	cm::acc
unggi	verb	verb::vv
fi	e	e::cve
toso	verb	verb::vv
ho	pst	e::pst.ptcp
,	mark	mark::comma

Examples of annotated POS tags for a Manchu sentence (Choi et al., 2023)

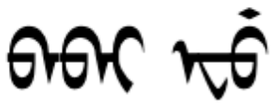
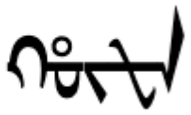
ManNER: Named Entity Recognition

Utilizing the “prpn (proper noun)” tags from the reference set,

We automatically annotate basic, personal, and place proper nouns.

Reference Tag	Our NER Tag	Meaning of the Tag
prpn	BSC	BaSiC proper noun
prpn.person	PER	PERsonal proper noun
prpn.plc	PLC	PLaCe proper noun

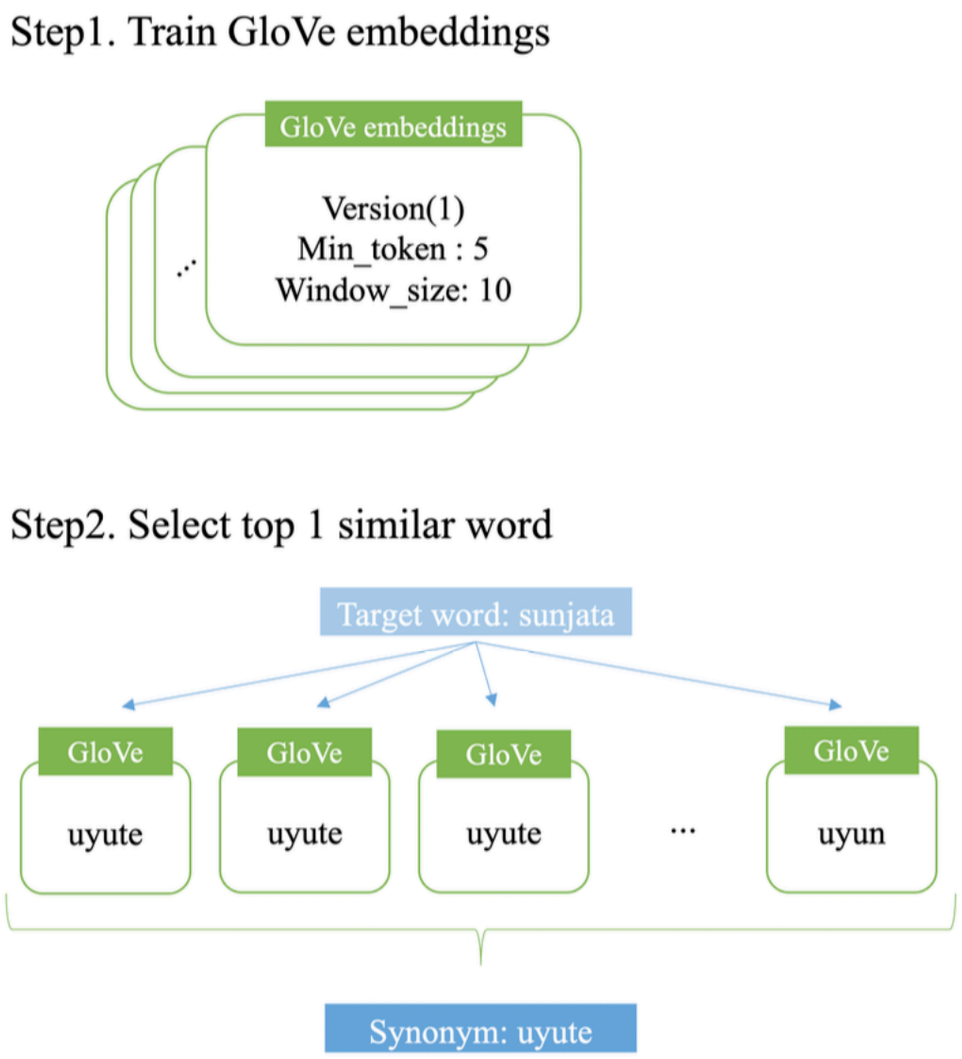
Examples of proper nouns

Manchu	Romanized form	tag
	babai efu	BSC
	bujantai	PER
	hoifa	PLC

Manchu Monolingual Corpus

Digitized, romanized monolingual Manchu data
→ for the purpose of training BERT models

Limitation of digitized Manchu text data → We need Data Augmentation!
(Seo et al., 2023)



Manchu Language Models

Based on the augmented monolingual Manchu corpus,

We train:

- Monolingual Manchu BERT
- mBERT adapted to Manchu
- BiLSTM-CRF

Monolingual Manchu BERT

Three versions of BERT models based on BERT-base config

Trained on:

- The entirety of accessible monolingual Manchu data without any augmentation
- Data with half augmentation
- Data with full augmentation

Each model is trained for 10 epochs, with the following hyperparameters.

- Vocabulary size of 25,000 (WordPiece-tokenized)
- Maximum sequence length of 512
- Training batch size of 10

mBERT Adapted to Manchu

Adapting the well-trained multilingual BERT to the unseen Manchu language

We expect the model to have learned general linguistic knowledge
and to additionally learn the properties of Manchu
and align them to its existing embedding space.

We continually pre-train the checkpoint of mBERT with our Manchu corpus.
(bert-multilingual-uncased)

- Training data: the ‘full augmentation’ version of monolingual Manchu texts
- Expand the mBERT vocab with 8,325 new tokens from Manchu
- Train the model for one epoch
- Following the configurations and settings of the original mBERT model
- Maximum sequence length of 512, training batch size of 16

BiLSTM-CRF for Specific Tasks

Trained for NER and two different versions of POS tagging, respectively:

- Trained on the ‘full augmentation’ version of monolingual Manchu texts
- Vocabulary size of 25,000
- Trained for 10 epochs
- Batch size of 32
- Max sequence length adjusted to accommodate the longest sequence in the training dataset
- Embedding dimension of 256
- Hidden dimension of 256
- Dropout rate of 0.5

Experiments

Split each dataset into training, validation, and test sets - 8:1:1

	Training Set	Validation Set	Test Set
# of examples	27,510	3,519	3,531

We perform each task experiment with six different models:

- BiLSTM (baseline)
  - Trained with the identical config to those of our BiLSTM-CRF model
  - Maximum sequence length of 50
- BiLSTM-CRF
- Manchu-adapted mBERT
- Three versions of pre-trained monolingual Manchu BERT LM with different portions of word replacement for data augmentation (no aug, half aug, full aug)

Experiments: NER

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

The half-augmented monolingual BERT shows the best F1 score.

The fully augmented BERT model shows slightly lower performance.

The baseline results are not very good:

→ discriminative capacity of our dataset (the complexity of the task)

Experiments: NER

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

Relatively low performance for personal proper nouns (PLC):

Due to the characteristics of personal proper nouns.

Many personal names in the Manchu language are derived from bare nouns.

e.g., “hashū” leaf side

Experiments: Simplified POS Tagging

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**

Nearly all of the models exhibit perfect performance.

- The number of simplified tags is smaller than that of the full tags.
  The simplified tags do not require consideration of additional information
  e.g. negation
- In Manchu language, the surface forms and POS almost exhibit
  a one-to-one correspondence.
  Manchu seldom incorporates irregular forms of inflection.

Experiments: Original POS Tagging

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

The mBERT-based model obtains the best performance.

No significant differences between BERT-based models and degrees of data augmentation

Slightly lower performance than that of simplified POS tagging

→ possibly due to the diverse and complicated POS tags

Experiments: Original POS Tagging

All the models exhibit low performance in tagging the classes:

- e::fve.imperative (imperative ending)

  Due to certain irregular forms of the imperative in the Manchu language

  e.g., daha- (to surrender) and baha- (to get) shares the same form

  for both imperative and past participle.

- e::fve.prv (preventive ending)
- e::fve.prs.prf.neg (negation of present perfect)

) low frequency

- e::npst.ptcp.neg (negation of past participle)

  Due to words that contain the substrings hakū and rakū (undurakü, cihakū)

  Misclassified, but the correct tags should be nv::nv and nv::adj

- nn::prpn.person (personal proper noun)

  Many personal names in the Manchu language are derived from bare nouns.

Analysis

Comparing the three versions of monolingual BERT models:

- The models trained on the augmented dataset report higher performance in all three tasks.
- The performance difference between half-aug and full-aug models is not very significant.

  → Increasing the ratio of word replacement

  → more copies of similar sentences with replaced words

  (while retaining the other words and overall syntax and semantics)

  → degrading the ability of the language model

  (harming the ability of LM to learn linguistic knowledge of Manchu)

Conclusion

We presented the first NLP task datasets for endangered Manchu language.

We constructed datasets for NER and POS tagging

based on the morphologically annotated Manchu corpus (Choi et al., 2023).

We trained the following language models as the task baselines:

The task-specific BiLSTM-CRF models, the Manchu-adapted multilingual BERT,

and three versions of monolingual Manchu BERT models.

Based on our augmented monolingual Manchu corpora.

Thank you for listening!

Sangah Lee, Sungjoo Byun, Jean Seo, Minha Kang

{sanalee, byunsj, seemdog, alsgk1123}@snu.ac.kr