NNVLP: A Neural Network-Based Vietnamese Language Processing Toolkit

Thai-Hoang Pham
Alt Inc
Hanoi, Vietnam
phamthaihoang.hn@gmail.com

Xuan-Khoai Pham
FPT University
Hanoi, Vietnam
khoaipxmse0060@fpt.edu.vn

Tuan-Anh Nguyen
Alt Inc
Hanoi, Vietnam
ntanh.hus@gmail.com

Phuong Le-Hong
Vietnam National University
Hanoi, Vietnam
phuonglh@vnu.edu.vn

Abstract

This paper demonstrates neural network-based toolkit namely NNVLP for essential Vietnamese language processing tasks including part-of-speech (POS) tagging, chunking, named entity recognition (NER). Our toolkit is a combination of bidirectional Long Short-Term Memory (Bi-LSTM), Convolutional Neural Network (CNN), Conditional Random Field (CRF), using pre-trained word embeddings as input, which achieves state-of-the-art results on these three tasks. We provide both API and web demo[1] for this toolkit.

1 Introduction

Vietnamese belongs to the top 20 most spoken languages and is employed by an important community all over the world. Therefore, research on Vietnamese language processing is an essential task. This paper focuses on three main tasks for Vietnamese language processing including POS tagging, chunking, and NER.

In this paper, we present a state-of-the-art system namely NNVLP for the Vietnamese language processing. NNVLP toolkit outperforms most previously published toolkits on three tasks including POS tagging, chunking, and NER. The contributions of this work consist of:

- We demonstrate a neural network-based system reaching the state-of-the-art performance for Vietnamese language processing including POS tagging, chunking, and NER. Our system is a combination of Bi-LSTM, CNN, and CRF models, which achieves an accuracy of 91.92%, F_1 scores of 84.11% and 92.91% for POS tagging, chunking, and NER tasks respectively.
- We provide our API and web demo for user, which is believed to positively contributing to the long-term advancement of Vietnamese language processing.

The remainder of this paper is structured as follows. Section 2 summarizes related work on Vietnamese language processing. Section 3 describes NNVLP toolkit architecture, API, and web interface. Section 4 gives experimental results and discussions. Finally, Section 5 concludes the paper.

2 Related Works

Previously published systems for Vietnamese language processing used traditional machine learning methods such as Conditional Random Field (CRF), Maximum Entropy Markov Model (MEMM), and Support Vector Machine (SVM). In particular, most of the toolkits for POS tagging task attempted to use conventional models such as CRF (Tran and Le, 2013) and MEMM (Le-Hong et al., 2010). (Tran and Le, 2013) also used CRF for chunking task. Recently, at the VLSP 2016 workshop for NER task, several participated system use MEMM (Le-Hong, 2016), (Nguyen et al., 2016) and CRF (Le et al., 2016) to solve this problem.

3 NNVLP API and Web Demo

3.1 System Architecture

We implement the deep neural network model described in (Pham and Le-Hong, 2017a). This

[1] nnvlp.org

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 37–40,
Taipei, Taiwan, November 27 – December 1, 2017.

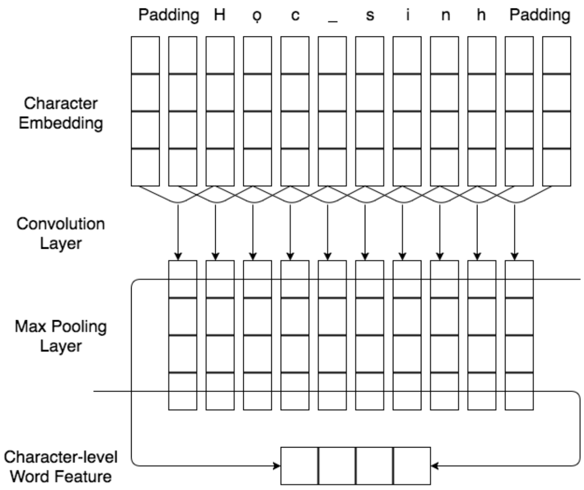

Figure 1: The CNN layer for extracting character-level word features of word *Học_sinh* (Student).

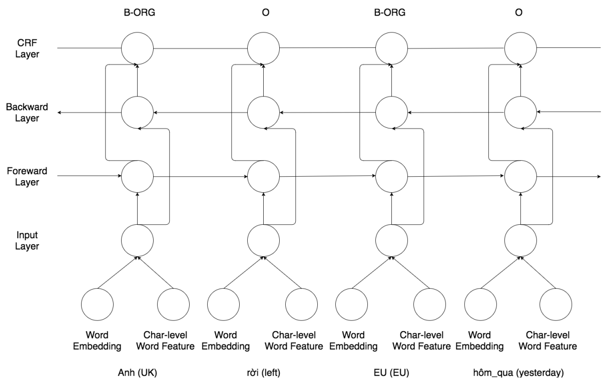

Figure 2: The Bi-LSTM-CRF layers for input sentence *Anh rời EU hôm qua.* (UK left EU yesterday.)

model is a combination of Bi-directional Long Short-Term Memory (Bi-LSTM), Convolutional Neural Network (CNN), and Conditional Random Field (CRF). In particular, this model takes as input a sequence of the concatenation of word embedding pre-trained by word2vec[2] tool and character-level word feature trained by CNN. That sequence is then passed to a Bi-LSTM, and then a CRF layer takes as input the output of the Bi-LSTM to predict the best named entity output sequence. Figure 1 and Figure 2 describe the architectures of BI-LSTM-CRF layers, and CNN layer respectively.

NNVLP toolkit uses these architectures for all tasks including POS tagging, chunking, and NER. Because each word in the Vietnamese language may consist of more than one syllables with spaces in between, which could be regarded as multiple words by the unsupervised models, we, first, segment the input texts into sequences of words by pyvi toolkit[3]. These word sequences are put into NNVLP toolkit to get corresponding POS tag sequences. Next, these words and POS tag sequences are put into NNVLP toolkit to get corresponding chunk sequences. Finally, NNVLP toolkit takes as input sequences of the concatenation of word, POS tag, and chunk to predict corresponding NER sequences. Figure 3 presents this pipeline of NNVLP toolkit.

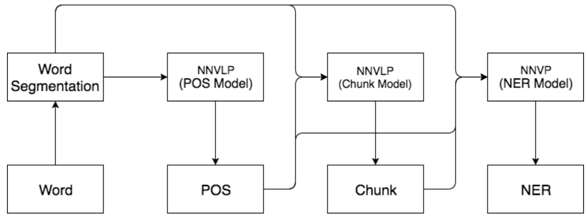

Figure 3: The Architecture of NNVLP Toolkit

3.2 NNVLP API

NNVLP API is an API for Vietnamese Language Processing which takes input sentences and outputs a JSON containing a list of sentences where each word in these sentences has POS tag, chunk, named entity attributes as shown in Figure 4.


```
{"sentences": [
  {
   "index": 0,
   "tokens": [
    {"chunk": "B-NP", "index": 0, "ner": "O", "pos": "Nc",
      "word": "Ông"},
    {"chunk": "I-NP", "index": 1, "ner": "B-PER", "pos": "Np",
      "word": "Nam"},
    {"chunk": "B-VP", "index": 2, "ner": "O", "pos": "V",
      "word": "là"},
    {"chunk": "B-NP", "index": 3, "ner": "O", "pos": "N",
      "word": "giảng_viên"},
    {"chunk": "B-NP", "index": 4, "ner": "B-ORG", "pos": "N",
      "word": "đại_học"},
    {"chunk": "I-NP", "index": 5, "ner": "I-ORG", "pos": "N",
      "word": "Bách_Khoa"},
    {"chunk": "O", "index": 6, "ner": "O", "pos": "CH", "word":
      "."}]
  }
 ]
}
```


Figure 4: The output JSON of the input sentence "Ông Nam là giảng viên đại học Bách Khoa." (Mr Nam is a lecturer of Bach Khoa University.)

[2] https://code.google.com/archive/p/word2vec/

[3] https://pypi.python.org/pypi/pyvi

3.3 Web Demo

We also provide web interface[4] for users of NNVLP toolkit. Users can type or paste raw texts into the textbox and click *Submit* button to get the coressponding POS tag, chunk, named entity sequences. Each label is tagged with different color to make the output easy to see. Users can also look up the meaning of each label by click *Help* button. Figure 5 presents the web interface of our system.

4 Experiments

In this section, we compare the performance of NNVLP toolkit with other published toolkits for Vietnamese including Vitk (Le-Hong et al., 2010), vTools (Tran and Le, 2013), RDRPOSTagger (Nguyen et al., 2014), and vie-ner-lstm (Pham and Le-Hong, 2017b).

4.1 Data Sets

To compare fairly, we train and evaluate these systems on the VLSP corpora. In particular, we conduct experiments on Viet Treebank corpus for POS tagging and chunking tasks, and on VLSP shared task 2016 corpus for NER task. All of these corpora are converted to CoNLL format. The corpus of POS tagging task consists of two columns namely word, and POS tag. For chunking task, there are three columns namely word, POS tag, and chunk in the corpus. The corpus of NER task consists of four columns. The order of these columns are word, POS tag, chunk, and named entity. While NER corpus has been separated into training and testing parts, the POS tagging and chunking data sets are not previously divided. For this reason, we use 80% of these data sets as a training set, and the remaining as a testing set. Because our system adopts early stopping method, we use 10% of these data sets from the training set as a development set when training NNVLP system. Table 1 and Table 2[5] shows the statistics of each corpus.

4.2 Evaluation Methods

We use the accuracy score that is the percentage of correct labels to evaluate the performance of each system for POS tagging task. For chunking and NER tasks, the performance is measured with F_1 score, where $F_1 = \frac{2*P*R}{P+R}$. Precision (P) is the percentage of chunks or named entities found by the learning system that are correct. Recall (R) is the percentage of chunks or named entities present in the corpus that are found by the system. A chunk or named entity is correct only if it is an exact match of the corresponding phrase in the data file.

[4] nnvlp.org

[5] For more details about these tagsets, please visit the demo website at nnvlp.org

	Number of sentences		
Data sets	POS	Chunk	NER
Train	7268	7283	14861
Dev	1038	1040	2000
Test	2077	2081	2831

Table 1: The number of sentences for each part in POS tagging, chunking, and NER data sets

Data sets	Labels
POS	N, V, CH, R, E, A, P, Np, M, C, Nc, L, T, Ny, Nu, X, B, S, I, Y, Vy
Chunk	NP, VP, PP, AP, QP, RP
NER	PER, LOC, ORG, MISC

Table 2: Labels in POS tagging, chunking, and NER data sets

4.3 Experiment Results

We evaluate performances of our system and several published systems on POS tagging, chunking, and NER data sets. Inputs for POS tagging task are words, for chunking task are words and POS tags, and for NER task are words, POS tags, and chunks. Table 3, Table 5, and Table 6 present the performance of each system on POS tagging, chunking, and NER task respectively. The hyper-parameters for training NNVLP are given in Table 4.

System	Accuracy
Vitk	88.41
vTools	90.73
RDRPOSTagger	91.96
NNVLP	**91.92**

Table 3: Performance of each system on POS tagging task

By combining Bi-directional Long Short-Term Memory, Convolutional Neural Network, and Conditional Random Field, our system outperforms most published systems on these three tasks. In particular, NNVLP toolkit achieves an accuracy of 91.92%, F_1 scores of 84.11% and 92.91% for

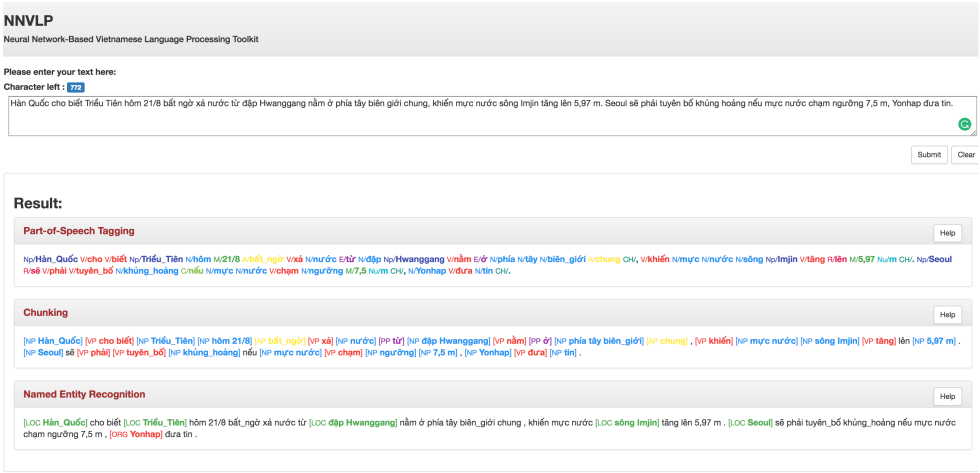

Figure 5: The Web Interface of NNVLP Toolkit

Layer	Hyper-parameter	Value
CNN	window size	3
	number of filters	30
LSTM	hidden nodes	300
Embedding	word	300
	character-level	30

Table 4: Hyper-parameters of our models

System	P	R	F1
vTools	82.79	83.55	83.17
NNVLP	83.93	84.28	**84.11**

Table 5: Performance of each system on chunking task

System	P	R	F1
Vitk	88.36	89.20	88.78
vie-ner-lstm	91.09	93.03	92.05
NNVLP	92.76	93.07	**92.91**

Table 6: Performance of each system on NER task

POS tagging, chunking, and NER tasks respectively.

5 Conclusion

We present a neural network-based toolkit for Vietnamese processing that is a combination of Bi-LSTM, CNN, and CRF. The system takes raw sentences as input and produces POS tag, chunk and named entity annotations for these sentences. The experimental results showed that NNVLP toolkit achieves state-of-the-art results on three tasks including POS tagging, chunking, and NER.

References

Thanh Huong Le, Thi Thu Trang Nguyen, Trong Huy Do, and Xuan Tung Nguyen. 2016. Named entity recognition in Vietnamese text. In *Proceedings of VLSP*, Hanoi, Vietnam.

Phuong Le-Hong. 2016. Vietnamese named entity recognition using token regular expressions and bidirectional inference. In *Proceedings of VLSP*, Hanoi, Vietnam.

Phuong Le-Hong, Azim Roussanaly, Thi Minh Huyen Nguyen, and Mathias Rossignol. 2010. An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In *TALN*, pages 50–61, Montreal, Canada.

Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. 2014. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In *Proceedings of the Demonstrations at EACL*, pages 17–20, Gothenburg, Sweden.

Thi Cam Van Nguyen, Thai Son Pham, Thi Hong Vuong, Ngoc Vu Nguyen, and Mai Vu Tran. 2016. Dsktlab-ner: Nested named entity recognition in Vietnamese text. In *Proceedings VLSP*, Hanoi, Vietnam.

Thai-Hoang Pham and Phuong Le-Hong. 2017a. End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character-level. In *Proceedings of PACLING*, pages 251–264, Yangon, Myanmar.

Thai-Hoang Pham and Phuong Le-Hong. 2017b. The importance of automatic syntactic features in Vietnamese named entity recognition. In *Proceedings of PACLIC*, Cebu, Philippines.

Mai-Vu Tran and Duc-Trong Le. 2013. vTools: Chunker and part-of-speech tools. *RIVF-VLSP 2013 Workshop*.