NLP@UIT at VLSP 2019: A Simple Ensemble Model for Vietnamese Dependency Parsing

Duc-Vu Nguyen, **Kiet Van Nguyen**, and **Ngan Luu-Thuy Nguyen**
University of Information Technology, Vietnam National University Ho Chi Minh City
{vund, kietnv, ngannlt}@uit.edu.vn

***Abstract*—This paper describes the system of the NLP@UIT research group that participated in Dependency Parsing for the Vietnamese task of the Vietnamese Language and Speech Processing 2019 shared-tasks. We developed a simple ensemble model for Vietnamese dependency parsing task. Our ensemble model uses two probability layers of the deep biaffine attention parser method with two different pre-trained word embeddings. According to the experimental result, our model achieved the promising effect on the development corpus. In our experiment, we were able to achieve the LAS score of 61.28 with the ensemble model on the private test dataset from VLSP 2019's organizers.**

***Index Terms*—Vietnamese; Natural Language Processing; Dependency Parsing**

I. Introduction

Dependency parsing is the task that generates grammatical relations between two words in the sentence. In recent years, dependency parsing is one of the topics that many researchers are interested in. For instance, the CoNLL 2017 [1] shared task and the CoNLL 2018 [2] shared task are the two most popular contests in the world for automatic dependency parsing on multiple languages.

The transition-based parsing and graph-based parsing are two models following data-driven parsers approach, which is learned from an annotated corpus and has flexible training methods than the rule-based systems [3], [4]. The last recent years have observed a rapid development of neural network methods. Many works proved that neural network methods could achieve state-of-the-art results on different tasks of natural language processing. For dependency parsing, we could see the same trend, many studies used deep neural networks to encode features without a hand-crafted representation [5]–[8]. Many ways for encoding the inputs for neural-network-based dependency parsing have proposed. Using pre-trained word embeddings is the most popular encoding method for several natural language processing tasks. Besides, the character-level word embeddings have been found useful for dependency parsing [9].

There are several studies on dependency parsing on Vietnamese. Thi *et al.* [10] and Nguyen *et al.* [11] converted the constituent treebank to the dependency treebank automatically. Nguyen and Nguyen [12] used supertags features on the transition-based parser. Vu-Manh *et al.* [13] used word embeddings on the transition-based parser. Nguyen *et al.* [14] used BiLSTMs encoder to generate word representation vectors and obtained a state-of-the-art result at publishing time on Vietnamese Dependency Treebank (VnDT) Nguyen *et al.* [11]. Specially, Nguyen [15] proposed the first joint multi-task model for Vietnamese word segmentation, part-of-speech (POS) tagging and dependency parsing. On benchmark Vietnamese datasets, experimental results show that their joint model obtains state-of-the-art or competitive performances. The study of Nguyen *et al.* [16] is one of rare Vietnamese natural language processing which uses the result of dependency parsing task for sentiment analysis task.

With the aim of Vietnamese natural language processing evolution, the organizers of Vietnamese Language and Speech Processing (VLSP) have proposed the new Vietnamese dependency treebank with the new annotation scheme. The new dependency treebank is used for dependency parsing for the Vietnamese task of VLSP 2019 shared-tasks[1]. The input of VLSP 2019 dependency parsing is the sentences that already contain gold word segmentation, universal part-of-speech tags, and Vietnamese part-of-speech tags. The task of VLSP 2019 dependency parsing is that predicts grammatical relations between two words in the sentence. According to the information of the VLSP 2019's organizers, the labeled attachment score (LAS) is the main metric of the VLSP 2019 dependency parsing task. We can see an dependency graph example of training dataset in Figure 1. We approach this task by building a simple ensemble model. Our ensemble model uses two probability layers of the deep biaffine attention parser method [17] with two different pre-trained word embeddings [18]. According to the experimental result, our model achieved a promising effect on our development dependency treebank.

The remainder of the paper is organized as follows. Section II describes our proposed system. The training details are represented in section III. We evaluate the results of the methods in section IV and draw the conclusion and future work in section V.

II. System Description

Our system at the VLSP 2019 dependency task entirely based on deep biaffine attention parser method [17]. Therefore, we briefly describe first deep biaffine attention parser method [17] in this section. After that, we present our simple ensemble method model using two probability layers of the deep biaffine attention parser method with two different pre-trained word embeddings.

A. Deep Biaffine Attention Parser

In this sub-section, we just briefly describe the dependency parser method of Stanford's team [19] at the CoNLL 2018 UD

[1] http://vlsp.org.vn/vlsp2019/eval/udp

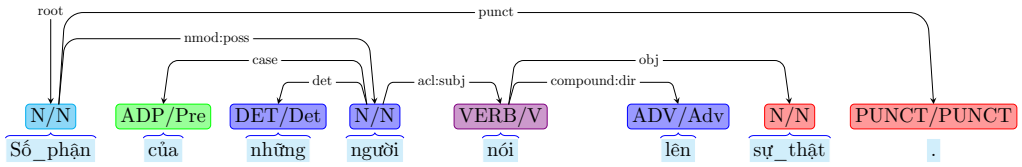

Figure 1. An dependency graph in training dataset of the VLSP 2019 dependency parsing task for Vietnamese, which contains gold word segmentation, universal part-of-speech tags, and Vietnamese part-of-speech tags.

Shared Task [2]. They have proposed few augmentations on the deep biaffine attention parser method [17].

Our dependency parser follows the study of Dozat and Manning [17], and Qi *et al.* [19]. We use the highway BiLSTM [20] for encoding the inputs, which are pretrained word embeddings, random word embeddings, character-level word embeddings, and Vietnamese part-of-speech tags (XPOS) embeddings. The unlabeled attachments are predicted by scoring each word i and its possible heads through a biaffine transformation [17]:

$$h_t = \text{BiLSTM}_t(x_1, \ldots, x_n), \quad (1)$$

$$v_i^{(ed)}, v_j^{(eh)} = \text{FC}^{(ed)}(h_i), \text{FC}^{(eh)}(h_j), \quad (2)$$

$$s_{ij}^{(e)} = [v_j^{(eh)}, 1]^\top U^{(e)} [v_i^{(ed)}, 1], \quad (3)$$

$$= \text{Deep-Biaff}^{(e)}(h_i, h_j), \quad (4)$$

$$P(y_{ij}^{(e)}|X) = \text{softmax}_j(s_i^{(e)}), \quad (5)$$

where $v_i^{(ed)}$ indicates for word i's edge-dependent representation, and $v_j^{(eh)}$ indicates for word i's edge-head representation. Qi *et al.* [19] have proposed the linearization models as an extensions of the deep biaffine attention parser [17]. They have factorized the relative location of word i and head j into their linear order and the distance between them with the conditional probability $P(y_{ij}|\text{sgn(i-j)}, \text{abs(i-j)})$, where $\text{sgn}(\cdot)$ is the sign function. The details of linearization models was presented in [19].

B. Ensemble Model

In the VLSP 2019 dependency parsing, the number of sentences in the training dataset is not large than 4,000, and the development dataset is not provided. Therefore, we decided to ensemble two probability layers of the deep biaffine attention parser method with two different pre-trained word embeddings. The final probability of unlabeled attachments is the sum from two models above with the α and $\alpha - 1$ coefficients. The final probability of dependent relations is the sum from two models above with the β and $\beta - 1$ coefficients. This ensemble is employed only at inference time.

III. Training Details

The dependency parser uses 3-layer BiLSTMs with 400d hidden states in forward and backward directions. We use 150d word embeddings, 150d character embeddings, and 150d Vietnamese POS tag embeddings. Pre-trained word embeddings and character-based word representations are both converted to be 150d. Throughout training, all embeddings are randomly replaced with a `<UNK>` symbol with p = .33 for learning unknown word/character/POS representations. The name of the first pre-trained embedding is `baomoi.model.bin`[2], whose dimension is 400d. This pre-trained word embedding is trained by the word2vec method with window-size 5. The name of the second pre-trained word embedding is `MULT_WC_F_E_B`[3] whose dimension is 2392d. This pre-trained word embedding is concatenated from multi other pre-trained word embeddings [18]. We apply dropout in all feed forward connections with p = .5 and use the training batch size 5000 [21]. Following Qi *et al.* [19], we use 400d fully connected layer with the ReLU non-linearity for both unlabeled attachments and dependent relations classifiers. We train the dependency parser system with RAdam [22] (α = .003, β_1 = .9, β_2 = .95) until 3,000 steps pass with no dev accuracy increases. Lastly, our implementation is based on the implementation of Qi *et al.* [21] with the source code available at: `https://stanfordnlp.github.io/stanfordnlp/`.

IV. Results

A. Results On The Development Dataset

In this section, we present the result of individual models when training with two turns of training dataset from the VLSP 2019's organizers.

TABLE I
Results On Randomly Development Dataset With Size Of 100 Sentences When Training Dataset Size Is 2,820 Sentences

Pre-trained Word Embedding	Stop Step	UAS	LAS
`MULT_WC_F_E_B` (2392d)	1,212	**81.77**	**72.12**
`baomoi.model.bin` (400d)	719	80.77	70.92

Table I shows that the deep biaffine attention parser method has the best LAS score (72.12) on the development dataset when using `MULT_WC_F_E_B` pre-trained word embedding.

TABLE II
Results On Public Test Dataset With Size Of 100 Sentences When Training Dataset Size Is 2,820 Sentences

Pre-trained Word Embedding	UAS	LAS
`MULT_WC_F_E_B` (2392d)	**80.01**	**69.21**
`baomoi.model.bin` (400d)	78.25	67.89

[2] https://github.com/sonvx/word2vecVN
[3] https://github.com/vietnlp/etnlp

Similarity to result presented in Table I, MULT_WC_F_E_B pre-trained word embedding also give the best LAS score (69.21) on public test dataset as we can see in Table II.

TABLE III
RESULTS OF ENSEMBLE MODELS ON PUBLIC TEST DATASET WITH SIZE OF 100 SENTENCES WHEN TRAINING DATASET SIZE IS 2,820 SENTENCES

α	β	**UAS**	**LAS**
0.50	0.50	80.46	70.24
0.73	0.66	**81.41**	**71.49**

After we evaluated deep biaffine attention parser method with two different pre-trained word embedding, we make an ensemble two results from two models with the method in sub-section II-B. We first evaluate the ensemble model with default coefficients $\alpha = 0.5$ and $\beta = 0.5$, in which α is the coefficient of the model using MULT_WC_F_E_B pre-trained word embedding and β is the coefficient of the model using baomoi.model.bin pre-trained word embedding. In Table III, we can see the ensemble model give better result than individual models with LAS score of 70.24. The turned coefficients $\alpha = 0.73$, $\beta = 0.66$ give the lower result than default coefficients $\alpha = 0.5$, $\beta = 0.5$ with LAS score of 71.49. The turned coefficients are different from default coefficients, which can be described by performances two models described in Table II.

TABLE IV
RESULTS ON RANDOMLY DEVELOPMENT DATASET WITH SIZE OF 100 SENTENCES WHEN TRAINING DATASET SIZE IS 3,856 SENTENCES

Pre-trained Word Embedding	**Stop Step**	**UAS**	**LAS**
MULT_WC_F_E_B (2392d)	1,845	79.08	70.07
baomoi.model.bin (400d)	3,314	**80.38**	**70.39**

Table IV shows that the deep biaffine attention parser method has the best LAS score (70.39) on the development dataset when using baomoi.model.bin pre-trained word embedding. The result of models when training dataset size is 3,956 sentences, which is different from the result in Table I. The reason can explain this result, which is the stylometry of the second turn dataset providing from VLSP 2019's organizers. Following the information from VLSP 2019's organizers, the stylometry of the first, and second turn dataset providing from VLSP 2019's organizers are *news* and, *literality*, respectively.

B. Results On The Private Dataset

TABLE V
RESULTS OF ENSEMBLE MODEL ($\alpha = 0.73$ AND $\beta = 0.63$) ON PRIVATE TEST DATASET (THE FIRST MODEL IS TRAINED BY 2,820 SENTENCES AND THE SECOND MODEL IS TRAINED BY 3,956 SENTENCES)

System	**News (400 sentences)**		**Literality (101 sentences)**		**Social (100 sentences)**		**Overall (601 sentences)**	
	UAS	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**
First Model	**72.95**	**60.19**	81.73	72.73	67.46	**57.08**	73.19	61.01
Second Model	72.85	60.08	**85.91**	**77.09**	**67.79**	56.75	**73.53**	**61.28**

Table V describes the official results of our system submitted to VLSP 2019's organizers on the private test dataset. We note that the our first model is trained by 2,920 sentences and our the second model is trained by 3,956 sentences. The difference between the stylometry from two training datasets of two models leads to the difference between the two results. The best LAS scores on two stylometry sentence groups *news* and *social* belong to the first model. However, the second model has obtained the highest LAS score overall with 61.28.

V. CONCLUSION & FUTURE WORKS

Our primary contribution to Dependency parsing for the Vietnamese task of VLSP 2019 shared-tasks is that using a simple ensemble model based on deep biaffine attention parser method. In our experiment, we were able to achieve the LAS score of 61.28 with the ensemble model on the private test dataset from VLSP 2019's organizers. Two of our submissions have proved that the larger amount of training data does not ensure the increase performances of the dependency parsers. We have to analyze the stylometry clusters of dependency treebank when we want to consider the increase performances of the dependency parsers. In the future works, we are planing two analyze the result of our system in more detail.

ACKNOWLEDGMENT

We would like to thank the VLSP 2019 organizers for their sustained hard work and providing the Vietnamese dependency treebank with the new scheme with high quality for this project. We would like to give our thanks to the NLP@UIT research group and the Multimedia Communications Laboratory of the University of Information Technology - Vietnam National University Ho Chi Minh City for their supports with pragmatic and inspiring advice.

REFERENCES

[1] D. Zeman, M. Popel, M. Straka, J. Hajic, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinkova, J. Hajic jr., J. Hlavacova, V. Kettnerová, Z. Uresova, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. dePaiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonca, T. Lando, R. Nitisaroj, and J. Li, "CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies", in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–19.

[2] D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies", in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, 2018, pp. 1–21.

[3] R. McDonald and J. Nivre, "Characterizing the errors of data-driven dependency parsing models", in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Association for Computational Linguistics, 2007, pp. 122–131.

[4] S. Buchholz and E. Marsi, "CoNLL-X Shared Task on Multilingual Dependency Parsing", in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, ser. CoNLL-X '06, Association for Computational Linguistics, 2006, pp. 149–164.

[5] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-Based Dependency Parsing with Stack Long Short-Term Memory", in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2015, pp. 334–343.

[6] M. Ballesteros, C. Dyer, Y. Goldberg, and N. A. Smith, "Greedy Transition-Based Dependency Parsing with Stack LSTMs", *Computational Linguistics*, vol. 43, no. 2, pp. 311–347, 2017.

[7] Kiperwasser, Eliyahu and Goldberg, Yoav, "Easy-First Dependency Parsing with Hierarchical Tree LSTMs", *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 445–461, 2016.

[8] E. Kiperwasser and Y. Goldberg, "Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations", *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 313–327, 2016.

[9] M. Ballesteros, C. Dyer, and N. A. Smith, "Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs", in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2015, pp. 349–359.

[10] L. N. Thi, L. H. My, H. N. Viet, H. N. T. Minh, and P. L. Hong, "Building a treebank for vietnamese dependency parsing", in *The 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, 2013, pp. 147–151.

[11] D. Q. Nguyen, D. Q. Nguyen, S. B. Pham, P.-T. Nguyen, and M. Le Nguyen, "From Treebank Conversion to Automatic Dependency Parsing for Vietnamese", in *Natural Language Processing and Information Systems*, E. Métais, M. Roche, and M. Teisseire, Eds., Springer International Publishing, 2014, pp. 196–207.

[12] K. V. Nguyen and N. L. Nguyen, "Vietnamese transition-based dependency parsing with supertag features", in *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*, 2016, pp. 175–180.

[13] C. Vu-Manh, A. T. Luong, and P. Le-Hong, "Improving Vietnamese Dependency Parsing Using Distributed Word Representations", in *Proceedings of the Sixth International Symposium on Information and Communication Technology*, ser. SoICT 2015, 2015, pp. 54–60.

[14] D. Q. Nguyen, M. Dras, and M. Johnson, "An empirical study for Vietnamese dependency parsing", *CoRR*, vol. abs/1611.00995, 2016. arXiv: 1611.00995.

[15] D. Q. Nguyen, "A neural joint model for Vietnamese word segmentation, POS tagging and dependency parsing", vol. abs/1812.11459, 2018. arXiv: 1812.11459.

[16] V. D. Nguyen, K. V. Nguyen, and N. L. Nguyen, "Variants of Long Short-Term Memory for Sentiment Analysis on Vietnamese Students' Feedback Corpus", in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Nov. 2018, pp. 306–311.

[17] T. Dozat and C. D. Manning, "Deep Biaffine Attention for Neural Dependency Parsing", *CoRR*, vol. abs/1611.01734, 2016. arXiv: 1611.01734.

[18] X.-S. Vu, T. Vu, S. N. Tran, and L. Jiang, "ETNLP: A Visual-Aided Systematic Approach to Select Pre-Trained Embeddings for a Downstream Task", in *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 2019.

[19] P. Qi, T. Dozat, Y. Zhang, and C. D. Manning, "Universal Dependency Parsing from Scratch", in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, 2018, pp. 160–170.

[20] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. R. Glass, "Highway Long Short-Term Memory RNNs for Distant Speech Recognition", *CoRR*, vol. abs/1510.08983, 2015. arXiv: 1510.08983.

[21] H. Zhou, Y. Zhang, S. Huang, and J. Chen, "A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing", in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2015, pp. 1213–1222.

[22] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the Variance of the Adaptive Learning Rate and Beyond", *arXiv e-prints*, 2019. arXiv: 1908.03265.