Đại học Quốc gia Thành phố Hồ Chí Minh

Trường Đại học Công nghệ Thông tin

ĐỒ ÁN CUỐI KỲ

Phương pháp luận Nghiên cứu Khoa học

CS519.O11

Họ và tên: Huỳnh Đặng Vĩnh Hiền

Lớp: KHTN2021

MSSV: 21520029

Giảng viên hướng dẫn: **PGS. TS. Lê Đình Duy**

Ngày 31 tháng 1 năm 2024

THÔNG TIN CHUNG CỦA NHÓM

- Đường dẫn video YouTube: `https://youtu.be/HcMIf8GTsw8`
- Đường dẫn slides: `https://github.com/vinhhien323/CS519`

Họ và tên: Huỳnh Đặng Vĩnh Hiền MSSV: 21520029 	Lớp: CS519.O11 Tự đánh giá (điểm tổng kết môn): 9.5 Số buổi vắng: 0 Số câu hỏi QT cá nhân: 13 Số câu hỏi QT cả nhóm: 13 (bao gồm 13 của cá nhân) Link Github: `https://github.com/vinhhien323`

ĐỀ CƯƠNG NGHIÊN CỨU

1 Tên đề tài (Tiếng Việt)

CẢI TIẾN HIỆU SUẤT BÀI TOÁN DEPENDENCY PARSING TIẾNG VIỆT DỰA TRÊN CƠ CHẾ ATTENTION KẾT HỢP ĐẶC TRƯNG NGÔN NGỮ

2 Tên đề tài (Tiếng Anh)

ENHANCING PERFORMANCE OF THE VIETNAMESE DEPENDENCY PARSING TASK BASED ON ATTENTION MECHANISM AND THE LINGUISTIC PATTERNS

3 Tóm tắt (Abstract)

(Tối đa 400 từ)

Our research is on Dependency Parsing, an NLP task related to the process of factorizing a natural language sentence into a tree-based structure called a dependency tree. This task has proven to have several applications both in NLP and real-life. The literature has shown that almost current research on the task takes into account the idea from Dozat and Manning [4], which uses the graph-based approach associated with encoder-decoder architecture and attention mechanism. On the other hand, the releasement of Transformers [16] has started a new race in model renovation. Several makeshift encoders like BERT [3], RoBERTa [8] or T5 [14] have raised the performance of NLP tasks so far. In fact, almost all NLP tasks, including Dependency Parsing, have a heavy dependence on the language. While the performance of the task is significantly high in English [12], it has still been a challenge in Vietnamese. To solve this task, we propose several approaches which keen on Vietnamese linguistic patterns, including but not limited to setting up a Vietnamese-based encoder (such as PhoBERT [9]), modifying the attention equation or using part-of-speech (POS) tagging. We have built several draft models, which show potential performance at 86.27% UAS and 78.43% LAS in the VnDT dataset [10], has also beaten part of the record of Vietnamese NLP tasks progress [17].

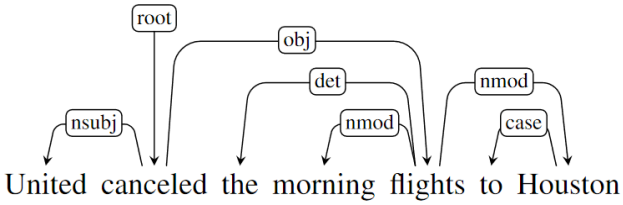
4 Giới thiệu (Introduction)

(Tối đa 1 trang A4)

Natural Language Processing, or NLP, is a subfield of Computer Science and Artificial Intelligence, focusing on solving computing tasks linked with human natural languages. In our research, we take the problem of Dependency Parsing, an NLP core task that focuses on the grammatical relationship between words in a sentence. In the task, a sentence written in natural language is mapped into a hierarchical structure called a dependency tree.

To build a dependency tree, in the initial stage, a word is chosen from the sentence to set as the *root* node. Then, each word is merged to the tree via an arc, linked from a word in the tree (called a *head*) to the considering word (called a *dependent*). Hence, an arc shows a binary grammatical

relationship. The process continues until every word in the sentence has been set, resulting in a directed tree.



An example of a dependency tree [6]

Moreover, each arc in a dependency tree is assigned a label to show the variant of the relationship, in other words, the effects of a head on its dependent and vice versa. A dataset in this task with human-annotated data is called a *tree bank*, where instances of dependency trees are given with a formalized set of labels.

The model that converts a sentence into its dependency tree, is called a *parser*. Currently, there are two main approaches to solving this problem consisting of transition-based and graph-based parsers [6]. In a transition-based parser, each word in the sentence is sequentially added to a *stack*, and whenever it is feasible, the top elements of the stack are dropped to create an arc. Recent research has taken into account the use of neural network and part-of-speech (POS) tagging in extracting the features and gaining scores for each action of the stack, which has archived significant performance in English [7]. The main benefit of transition-based approaches is low computational complexity but face difficulties with long sentences or languages with loose grammatical structure.

On the other hand, the graph-based parser gets the idea from graph theory. The parser starts with assigning a score for every pair of words in the sentence, then uses the Chu-Liu [1] Edmonds [5] algorithm to select the maximum spanning tree. Almost all graph-based research has chosen to keep the algorithm of the approach and tried to find an effective method to set up the arc scores. The advantages of this approach are its intuitive and the flexibility to deal with almost all languages, with the trace off of computing performance. In fact, the paper of Dozat and Manning [4] has a great impact on the research society. In their solution, a sentence is firstly processed via an encoder (such as LTSM) to collect its representation, which is later an input for the biaffine attention classifier. Finally, the result of the attention equation is passed by several MLPs to receive the arc scores.

Several inherited models from Dozat and Manning [4] can be represented such as:

- [2], where auxiliary week models are built to take advantage of unlabeled data. Moreover, this paper proposed a multi-task training solution with competitive performance.
- [19], where the LSTM is replaced with next-generation encoders including BERT and RoBERTa. Hence, the model increases its flexibility with languages and also enhances its performance.
- [15], where the authors build up a vast auto-generated tree bank from a pre-released parser, and then use it as data for the pre-training step. At the fine-tuning step, the gold-labeled dataset is used as usual. In another view, this can be seen as a *cross-domain* model due to the difference between the dataset in two steps.

For the Dependency Parsing task in Vietnamese, there are only a few number of publications. From the Vietnamese NLP tasks progress [17], we have solutions from [9] and [11], where the authors build a new encoder to deal with Vietnamese (named PhoBERT) and use it to train multiple NLP tasks, including POS and name entity recognition (NER).

5 Mục tiêu (Goals)

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)
We are keen on building a parser to adapt at least one of two criteria:

- Able to beat the record on [17] to make the new state-of-the-art model in this task.
- Able to handle the cross-domain problem between different datasets.

6 Nội dung và Phương pháp (Methodology)

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)
Firstly, we want to implement the research of [7] in Vietnamese. In addition, we think it is feasible to add other linguistic patterns including NER because the proper name in a sentence usually plays an important role in the meaning structure of a sentence. On the other hand, consistent-based dependency parsing in Vietnamese is extremely rare, so we think it is interesting and potential to try it. We hope this approach can help to improve the performance of models.

On the other hand, we consider taking unlabeled and, machine-generated data into account such as [2] [15] because Vietnamese is known to have a great resource of data, especially on the Internet. Moreover, this approach helps the model to learn the general meaning of language in various scenarios. However, a large amount of data requires a high-performance computing system, which may be costly.

As there are several different dependency treebanks in Vietnamese, we want to solve the cross-domain problem. Taking the idea of modifying the biaffine attention from [15], we want to apply the same methodology to Vietnamese.

7 Kết quả mong đợi (Expected Result)

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)
From our goals, we have set several excepted quantitative results as follows:

- For the goal of building a high-performance parsing model, in our perspective, the result of about 90% unlabeled attachment score (UAS) and/or 85% labeled attachment score (LAS) in the VnDT dataset are desirable. These will create a margin of 5% of both evaluation metrics, compared to current state-of-the-art solutions.
- For the task of cross-domain dependency parsing, it is difficult to evaluate due to the fact that there are not any baselines for this task in Vietnamese. From research of [18] and [13], a decrease of 10% in performance contrasting to a non-cross-domain task is reasonable, resulting in 70% UAS and 75% LAS.

8 Tài liệu tham khảo (References)

(Định dạng DBLP)

[1] Y. CHU. "On the shortest arborescence of a directed graph". In: *Science Sinica* 14 (1965), pp. 1396–1400. URL: https://cir.nii.ac.jp/crid/1570854175817997952.

[2] Kevin Clark et al. *Semi-Supervised Sequence Modeling with Cross-View Training*. 2018. arXiv: 1809.08370 [cs.CL].

[3] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[4] Timothy Dozat and Christopher D. Manning. *Deep Biaffine Attention for Neural Dependency Parsing*. 2017. arXiv: 1611.01734 [cs.CL].

[5] Jack Edmonds et al. "Optimum branchings". In: *Journal of Research of the national Bureau of Standards B* 71.4 (1967), pp. 233–240.

[6] Daniel Jurafsky and James H Martin. *Speech and Language Processing*. 3rd (draft). Redwood, California, United States: Stanford University Press, 2024.

[7] Eliyahu Kiperwasser and Yoav Goldberg. *Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations*. 2016. arXiv: 1603.04351 [cs.CL].

[8] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[9] Dat Quoc Nguyen and Anh Tuan Nguyen. *PhoBERT: Pre-trained language models for Vietnamese*. 2020. arXiv: 2003.00744 [cs.CL].

[10] Dat Quoc Nguyen et al. "From Treebank Conversion to Automatic Dependency Parsing for Vietnamese". In: June 2014, pp. 196–207. ISBN: 978-3-319-07982-0. DOI: 10.1007/978-3-319-07983-7_26.

[11] Linh The Nguyen and Dat Quoc Nguyen. *PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing*. 2021. arXiv: 2101.01476 [cs.CL].

[12] *Papers with Code - Dependency Parsing — paperswithcode.com*. https://paperswithcode.com/task/dependency-parsing. [Accessed 28-01-2024].

[13] Xue Peng et al. "Overview of the NLPCC 2019 Shared Task: Cross-Domain Dependency Parsing". In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*. Dunhuang, China: Springer-Verlag, 2019, pp. 760–771. ISBN: 978-3-030-32235-9. DOI: 10.1007/978-3-030-32236-6_69. URL: https://doi.org/10.1007/978-3-030-32236-6_69.

[14] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2023. arXiv: 1910.10683 [cs.LG].

[15] Yuanhe Tian, Yan Song, and Fei Xia. "Enhancing Structure-aware Encoder with Extremely Limited Data for Graph-based Dependency Parsing". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 5438–5449. URL: https://aclanthology.org/2022.coling-1.483.

[16] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[17] *Vietnamese NLP tasks — nlpprogress.com*. https://nlpprogress.com/vietnamese/vietnamese.html. [Accessed 28-01-2024].

[18] Yi Zhang and Rui Wang. "Cross-domain dependency parsing using a deep linguistic grammar". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*. ACL '09. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 378–386. ISBN: 9781932432459.

[19] Yu Zhang, Zhenghua Li, and Min Zhang. *Efficient Second-Order TreeCRF for Neural Dependency Parsing*. 2020. arXiv: 2005.00975 [cs.CL].