ENHANCING PERFORMANCE OF THE VIETNAMESE DEPENDENCY PARSING TASK

based on attention mechanism and the linguistic patterns

Huynh Dang Vinh Hien, 21520029

Falcuty of Computer Science,
University of Information Technology,
Vietnam National Univeristy - Ho Chi Minh City

January 31, 2024

Table of Contents

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Research Introduction

Research Introduction

- Natural Language Processing, or NLP, is a field of research about how to simulate and solve lingual-related problems on computers.

Research Introduction

- Natural Language Processing, or NLP, is a field of research about how to simulate and solve lingual-related problems on computers.
- Has several applications such as machine translation system or virtual assistance, with contribute to enhance people's lives.

Research Introduction

- Natural Language Processing, or NLP, is a field of research about how to simulate and solve lingual-related problems on computers.
- Has several applications such as machine translation system or virtual assistance, with contribute to enhance people's lives.
- A particular of NLP tasks is they depends on the natural language used in the models.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Research Introduction

- Natural Language Processing, or NLP, is a field of research about how to simulate and solve lingual-related problems on computers.
- Has several applications such as machine translation system or virtual assistance, with contribute to enhance people's lives.
- A particular of NLP tasks is they depends on the natural language used in the models.
- Some tasks has been solved totally in English, but have not seen significantly effective models in Vietnamese.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

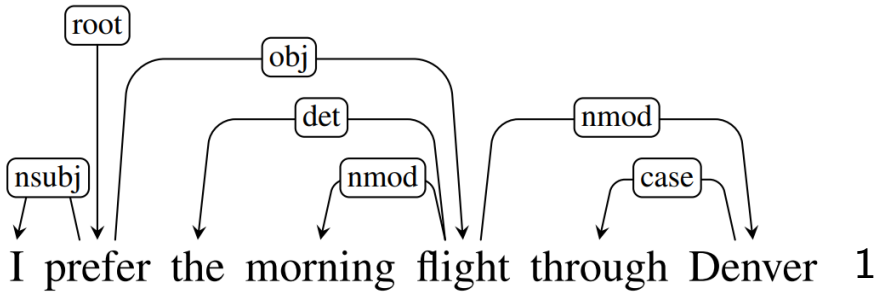
Research Introduction

- Natural Language Processing, or NLP, is a field of research about how to simulate and solve lingual-related problems on computers.
- Has several applications such as machine translation system or virtual assistance, with contribute to enhance people's lives.
- A particular of NLP tasks is they depends on the natural language used in the models.
- Some tasks has been solved totally in English, but have not seen significantly effective models in Vietnamese.
- Including our researching task: **Dependency Parsing**.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Dependency Grammar

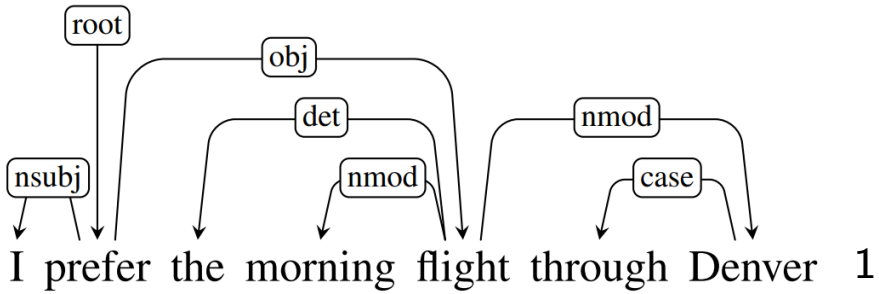
- Dependency grammar is a grammar formalism, where each word of a sentence is describe by a binary relationship called **head** and **dependent**, which is used to build a **dependency tree**.



[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Dependency Grammar

- Dependency grammar is a grammar formalism, where each word of a sentence is describe by a binary relationship called **head** and **dependent**, which is used to build a **dependency tree**.



- Each arc in the given picture illustrates a grammatical relation from a head to a dependent and comes with a label to classify the relation.

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Why Dependency Grammar?

- In addition, we already have context-free grammar and constituent-based representations.

Why Dependency Grammar?

- In addition, we already have context-free grammar and constituent-based representations.
- Dependency Grammar has performed better in case of long sentences.

Why Dependency Grammar?

- In addition, we already have context-free grammar and constituent-based representations.
- Dependency Grammar has performed better in case of long sentences.
- Some languages does not have strict grammatical rules, with makes the process of creating rules become difficult.

Why Dependency Grammar?

- In addition, we already have context-free grammar and constituent-based representations.
- Dependency Grammar has performed better in case of long sentences.
- Some languages does not have strict grammatical rules, with makes the process of creating rules become difficult.
- For example, in Vietnamese, adjective can come both before and after the noun, depends on the scenario.

Hồ Tây

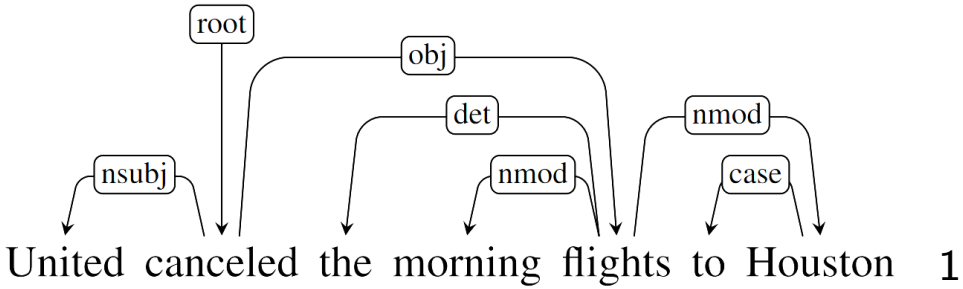
or

Tây Hồ

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Dependency Relations and Dependency Treebanks

- which means the labels of dependency arcs.

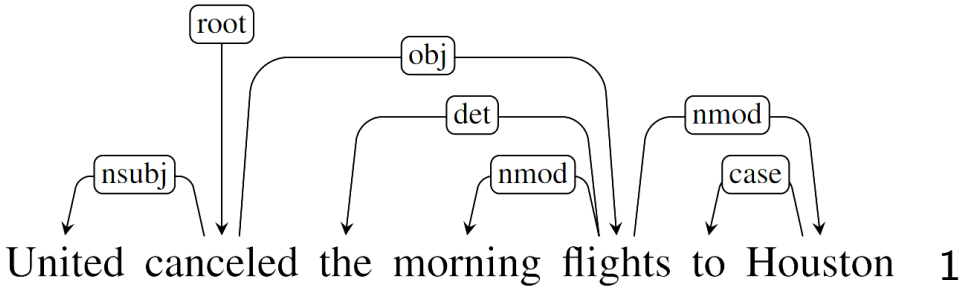


[1]

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Dependency Relations and Dependency Treebanks

- which means the labels of dependency arcs.



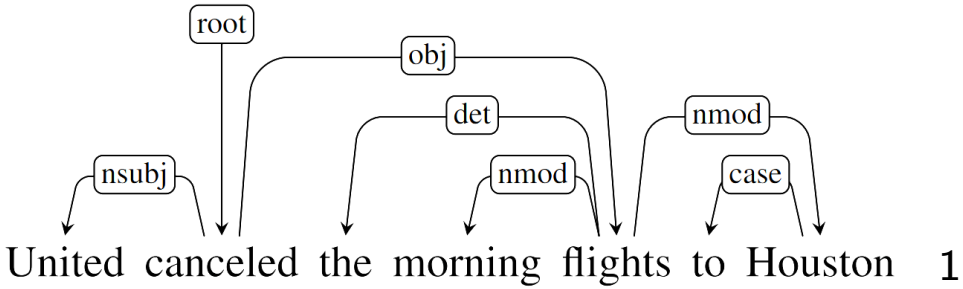
[1]

- In this example, the arc links between *canceled* and *flights* is marked as *obj*, which mean *object*.

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Dependency Relations and Dependency Treebanks

- which means the labels of dependency arcs.



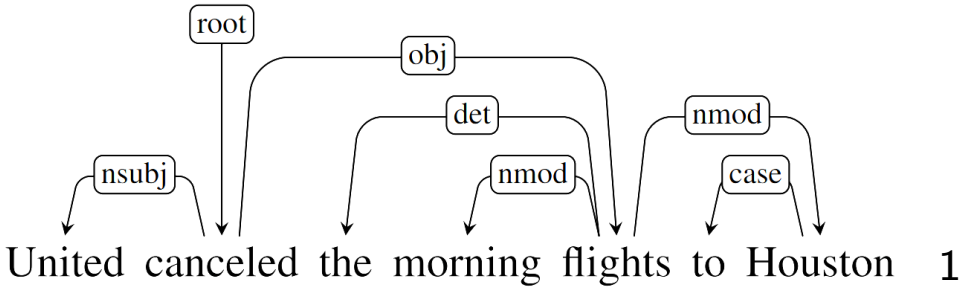
[1]

- In this example, the arc links between *canceled* and *flights* is marked as *obj*, which mean *object*.
- Some projects, like **Universal Dependencies**, have released formality definitions of dependency relations.

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Dependency Relations and Dependency Treebanks

- which means the labels of dependency arcs.



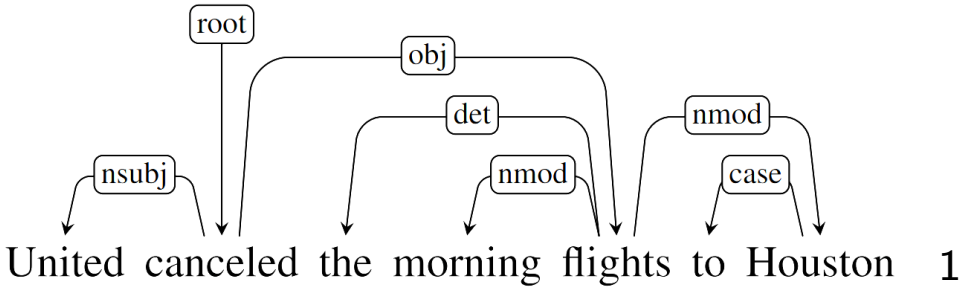
[1]

- In this example, the arc links between *canceled* and *flights* is marked as *obj*, which mean *object*.
- Some projects, like **Universal Dependencies**, have released formality definitions of dependency relations.
- A dependency parsing dataset with gold labels is called a *treebank*.

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Dependency Relations and Dependency Treebanks

- which means the labels of dependency arcs.



[1]

- In this example, the arc links between *canceled* and *flights* is marked as *obj*, which mean *object*.
- Some projects, like **Universal Dependencies**, have released formality definitions of dependency relations.
- A dependency parsing dataset with gold labels is called a *treebank*.
- As there is a gap between treebanks, which can cause a *cross-domain* problem.

[1] *Speech and Language Processing* by Daniel Jurafsky and James H. Martin

Transition-based Dependency Parsing

- Takes the idea from the *stack* data structure.

Transition-based Dependency Parsing

- Takes the idea from the *stack* data structure.
- The words in a sentence is sequentially put into the stack.

Transition-based Dependency Parsing

- Takes the idea from the *stack* data structure.
- The words in a sentence is sequentially put into the stack.
- When ever it is reasonable, the top two elements from the stack are pop out to create an arc.

Transition-based Dependency Parsing

- Takes the idea from the *stack* data structure.
- The words in a sentence is sequentially put into the stack.
- When ever it is reasonable, the top two elements from the stack are pop out to create an arc.
- Current research focuses on using machine learning and neural network to extract feature and gain score for each transition.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Graph-based Dependency Parsing

- This method takes idea from graph theory.

[2] Zhuoran Wang et al. 2023, Efficient and Effective Directed Minimum Spanning Tree Queries

Graph-based Dependency Parsing

- This method takes idea from graph theory.
- Suppose that we have a sentence consists of n words, we will transform it into a n vertices fully connected directed graph.

[2] Zhuoran Wang et al. 2023, Efficient and Effective Directed Minimum Spanning Tree Queries

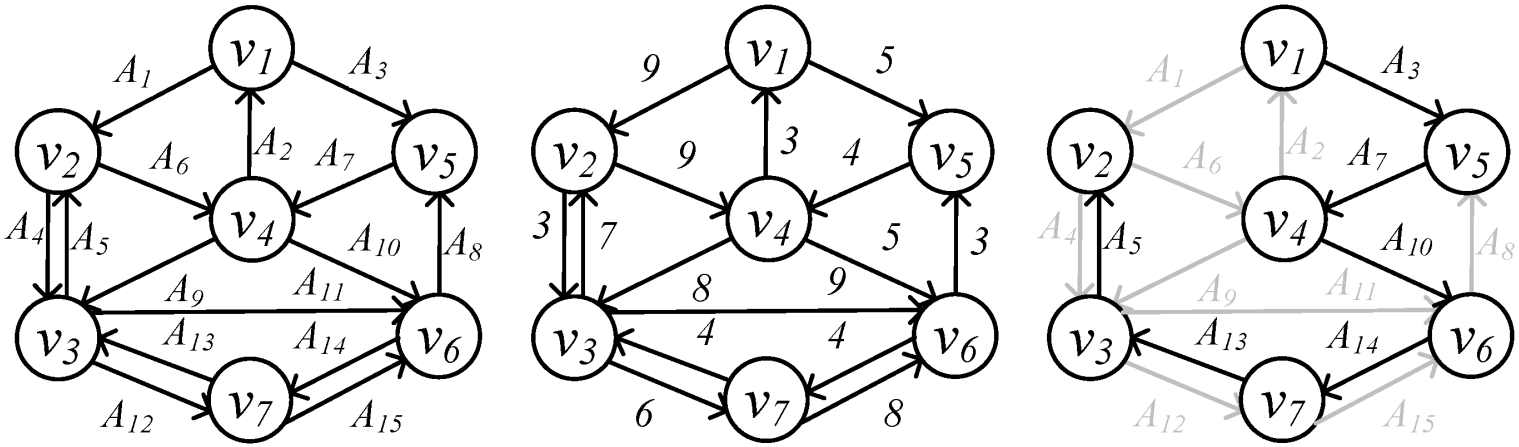
Graph-based Dependency Parsing

- This method takes idea from graph theory.
- Suppose that we have a sentence consists of n words, we will transform it into a n vertices fully connected directed graph.
- Then, each of $n(n-1)$ edges in this graph is gained a score.

[2] Zhuoran Wang et al. 2023, Efficient and Effective Directed Minimum Spanning Tree Queries

Graph-based Dependency Parsing

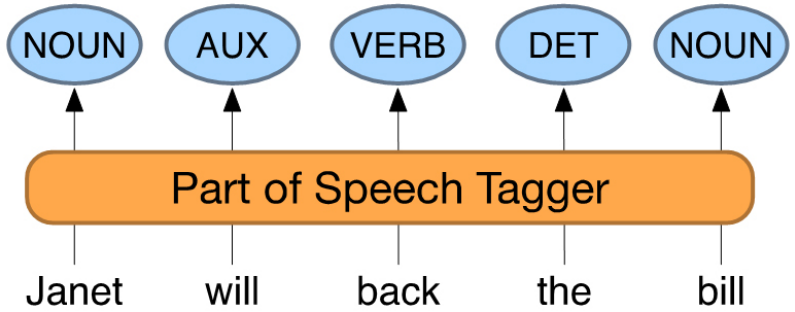
- This method takes idea from graph theory.
- Suppose that we have a sentence consists of n words, we will transform it into a n vertices fully connected directed graph.
- Then, each of $n(n-1)$ edges in this graph is gained a score.
- Use Chu-Liu Edmonds (a variance of MST algorithm) to build the dependency tree.



[2] Zhuoran Wang et al. 2023, Efficient and Effective Directed Minimum Spanning Tree Queries

Linguistic Patterns

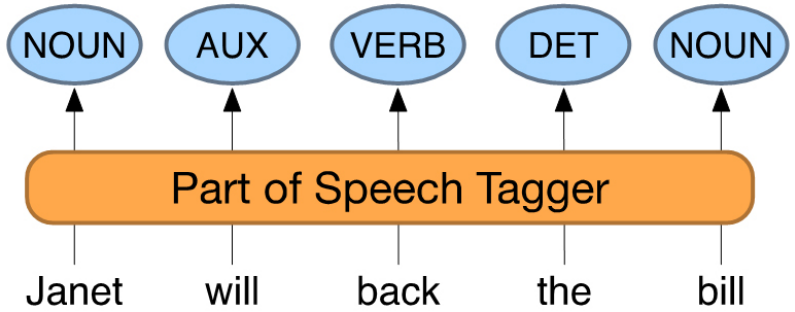
- There are several linguistic patterns in NLP including part-of-speech (POS) tagging, name entity recognition (NER) or correference resolution



An example of part-of-speech (POS) tagging

Linguistic Patterns

- There are several linguistic patterns in NLP including part-of-speech (POS) tagging, name entity recognition (NER) or correference resolution

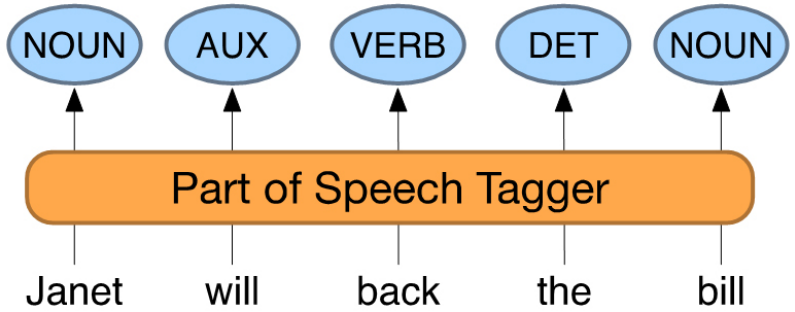


An example of part-of-speech (POS) tagging

- Some research has shown that adding these patterns as an input to NLPs may help to receive better performance.

Linguistic Patterns

- There are several linguistic patterns in NLP including part-of-speech (POS) tagging, name entity recognition (NER) or correference resolution



An example of part-of-speech (POS) tagging

- Some research has shown that adding these patterns as an input to NLPs may help to receive better performance.
- We want to try applying this approach for the task in Vietnamese, as it has flexible word orders.

Research interests

Research interests

- Our goal is to solve the Vietnamese Dependency Parsing task that satisfies at least one in two criteria:

Research interests

- Our goal is to solve the Vietnamese Dependency Parsing task that satisfies at least one in two criteria:
  - Able to beat the record on to make the new state-of-the-art model in this task.

Research interests

- Our goal is to solve the Vietnamese Dependency Parsing task that satisfies at least one in two criteria:
  - Able to beat the record on to make the new state-of-the-art model in this task.
  - Able to handle the cross-domain problem between different datasets.

Research interests

- Our goal is to solve the Vietnamese Dependency Parsing task that satisfies at least one in two criteria:
  - Able to beat the record on to make the new state-of-the-art model in this task.
  - Able to handle the cross-domain problem between different datasets.
- Our approach is based on current research, then applying next-generation encoders (such as BERT, RoBERTa) and adding Vietnamese linguistic patterns.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Methodology

Methodology: Enhancing performance

- We propose to implement the parser of (Eliyahu Kiperwasser and Yoav Goldberg, 2017) in Vietnamese there are only a few of research about transition-based parser in Vietnamese.

Methodology: Enhancing performance

- We propose to implement the parser of (Eliyahu Kiperwasser and Yoav Goldberg, 2017) in Vietnamese there are only a few of research about transition-based parser in Vietnamese.
- As almost all current graph-based parsers take the idea of (Dozat and Manning, 2017), we want to reimplement it by replacing the LSTM with next-generation encoders.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Methodology: Enhancing performance

- We propose to implement the parser of (Eliyahu Kiperwasser and Yoav Goldberg, 2017) in Vietnamese there are only a few of research about transition-based parser in Vietnamese.
- As almost all current graph-based parsers take the idea of (Dozat and Manning, 2017), we want to reimplement it by replacing the LSTM with next-generation encoders.
- Another point of view is taking the unlabeled data and multi-task training of (Clark et al., 2018), but this solution require dataset for various tasks.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Methodology: Cross-domain problem

- As there are limited research on the cross-domain problem of this task, we think this view may be feasible to experiment.

Methodology: Cross-domain problem

- As there are limited research on the cross-domain problem of this task, we think this view may be feasible to experiment.
- (Tian et al., 2022) is an instance where the biaffine attention is adapted to face with multiple domains. However, this solution use a large amount of data, we could only try this view in case of having high-performance system.

Preliminary Results

Preliminary Results

We have implemented the solution of (Tian et al., 2022) and (Zhang et al, 2020) with the VnDT treebank by replacing the lstm encoder with PhoBERT and recevied the following result.

Model	UAS	LAS
(Tian et al., 2022)	85.66%	77.88%
(Zhang et al, 2020)	85.60%	77.24%

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN