Tổng quát hoá miền tri thức trên bài toán phân tích cú pháp phụ thuộc tiếng Việt

SVTH: Huỳnh Đặng Vĩnh Hiền, Lê Châu Anh
GVHD: TS. Nguyễn Thị Quý, ThS. Huỳnh Thiện Ý

Khoa Khoa học Máy tính,
Trường Đại học Công nghệ Thông tin,
Đại học Quốc gia Thành phố Hồ Chí Minh

Ngày 15 tháng 1 năm 2025

Mục lục

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

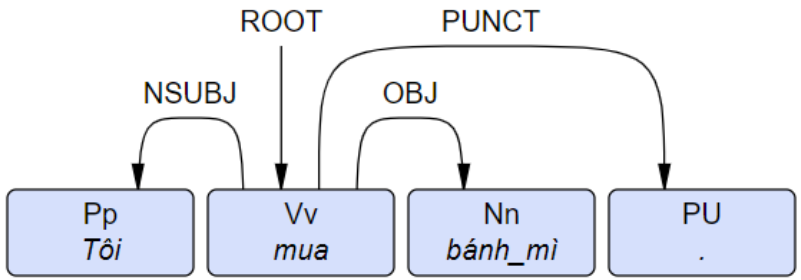
Tổng quan khoá luận

Bài toán phân tích cú pháp phụ thuộc

Định nghĩa

Phân tích cú pháp phụ thuộc là một bài toán trong Xử lý ngôn ngữ tự nhiên, tập trung vào phân tích cấu trúc ngữ pháp, cụ thể là mối quan hệ phụ thuộc giữa các từ trong câu.

Để thể hiện cấu trúc ngữ pháp, mỗi câu sẽ được biểu diễn dưới dạng một cây phụ thuộc.



Hình: Một ví dụ về cây phụ thuộc

Kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt

Trong bài toán phân tích cú pháp phụ thuộc, tiếng Việt vẫn được xem là một ngôn ngữ có tài nguyên hạn chế.

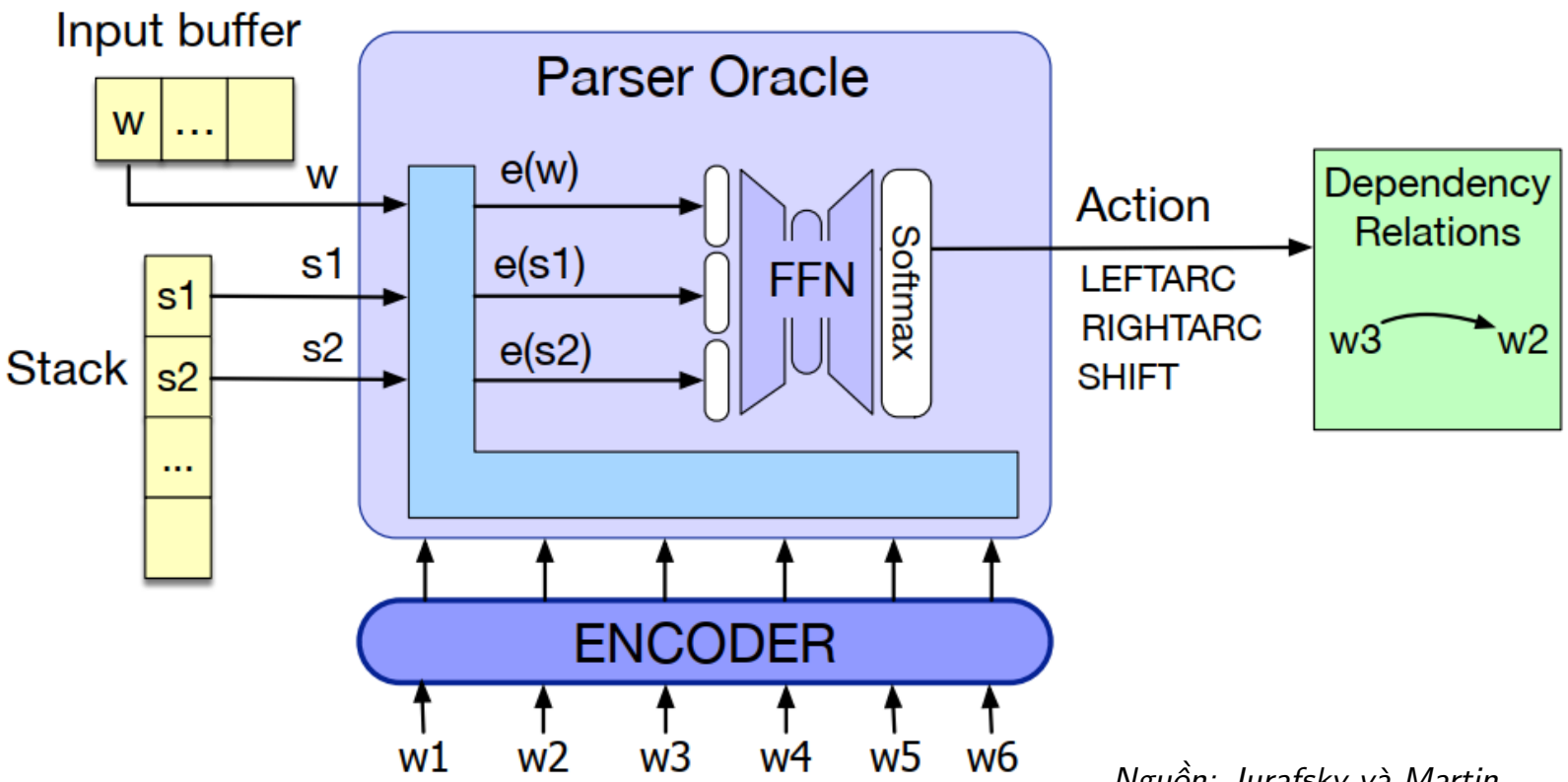
Bảng: Tổng quan các kho ngữ liệu cú pháp phụ thuộc tiếng Việt

Kho ngữ liệu	**Số câu**	**Số nhãn phụ thuộc**	**Phương pháp gán nhãn**	**Nguồn ngữ liệu**
BKTreebank	6,909	26	Thủ công	Báo Dân Trí
VnDT	10.197	33	Tự động	Báo Tuổi Trẻ
UD_Vietnamese-VTB	3.323	84	Tự động	Báo Tuổi Trẻ

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Phân tích cú pháp phụ thuộc dựa trên chuyển đổi (transition-based)

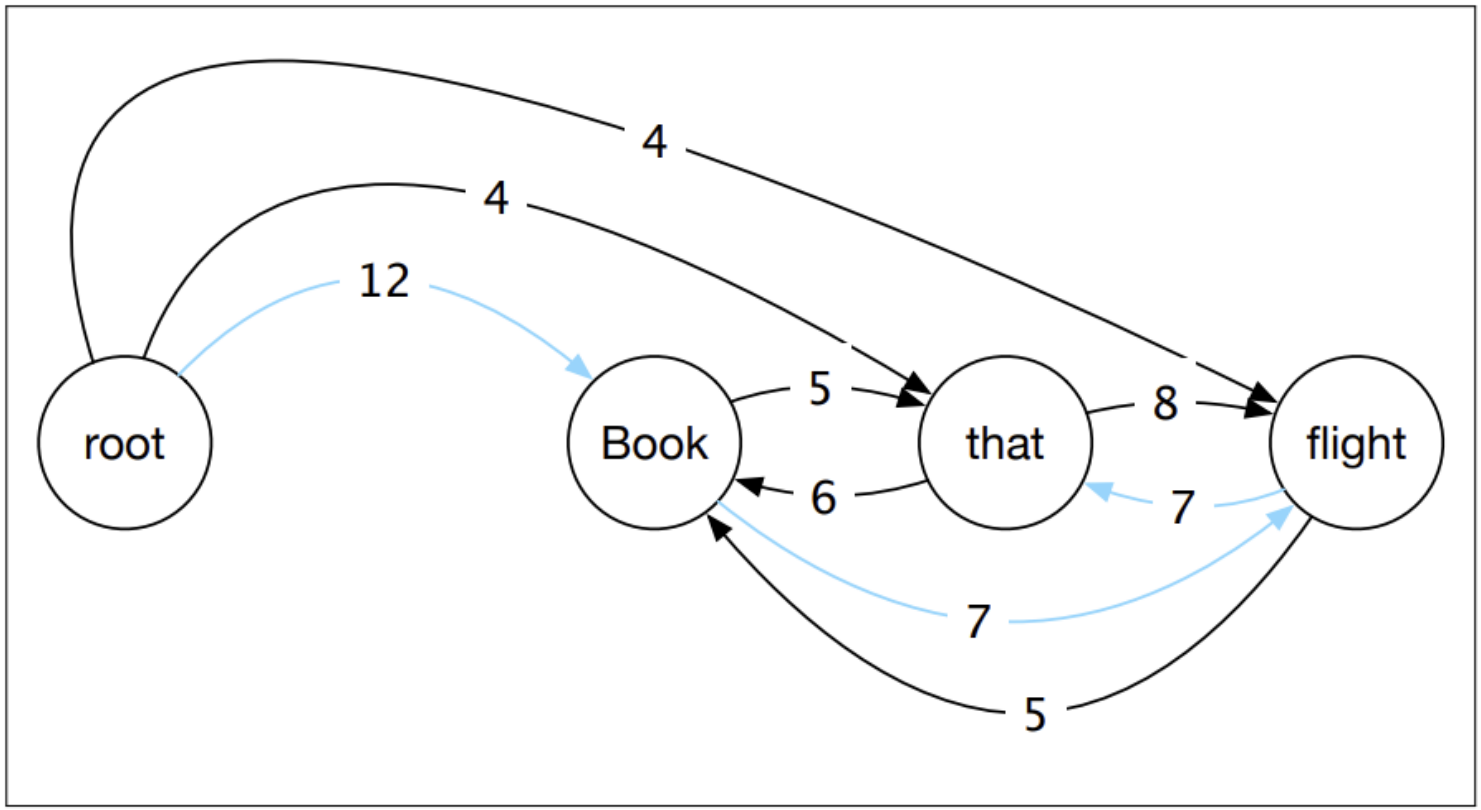
Ý tưởng: Xây dựng cấu trúc phụ thuộc bằng cách sử dụng các toán tử dịch chuyển khi duyệt câu một cách tuần tự.



Nguồn: Jurafsky và Martin

Phân tích cú pháp phụ thuộc dựa trên đồ thị (graph-based)

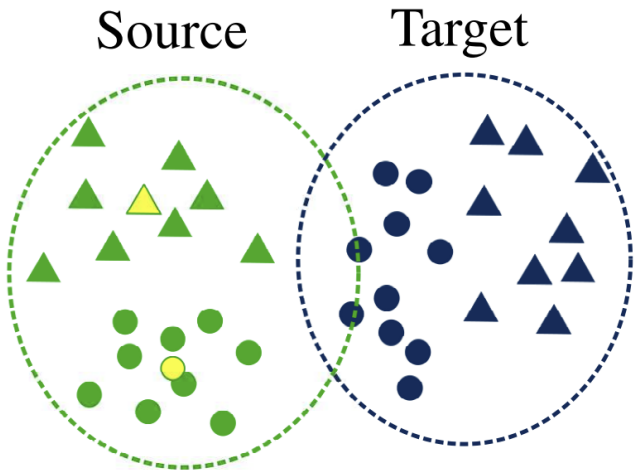
Ý tưởng: Mô hình hóa việc tìm cấu trúc cú pháp tối ưu cho câu thành bài toán tìm cây khung lớn nhất.



Nguồn: Jurafsky và Martin

Vấn đề khác biệt miền tri thức

- Hầu hết nghiên cứu về phân tích cú pháp phụ thuộc không xem xét đến bối cảnh có sự khác biệt về phân phối giữa các miền tri thức (domain gap).
- Tuy nhiên, đây là một thách thức phổ biến khi triển khai mô hình với dữ liệu thực tế, gây khó khăn và làm giảm hiệu suất của mô hình.
- Chỉ có một số ít nghiên cứu liên quan tới domain gap nhưng chủ yếu mang tính khảo sát và chưa nghiên cứu trên tiếng Việt.



Nguồn: Robert và cộng sự

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Miền tri thức

Định nghĩa

Miền tri thức là một tập hợp dữ liệu có chung ngữ cảnh và các quy tắc sử dụng ngôn ngữ.

Miền tri thức có thể được phân loại dựa trên các yếu tố như chủ đề, phong cách viết, hoặc mức độ trang trọng . . .

Nếu mỗi chủ đề được coi là một miền tri thức, đặc điểm nổi bật giữa các miền tri thức là sự khác biệt về kho từ vựng chuyên ngành.

Miền tri thức

Miền tri thức *Kinh doanh*:

- Dù chứng khoán đảo chiều vào cuối phiên để tăng 4 điểm, thanh khoản vẫn rơi về mức thấp nhất kể từ cuối tháng 10/2023.
- Theo Chứng khoán Vietcombank (VCBS), thanh khoản sụt giảm cho thấy lực cung bán ra đã có phần chững lại.

Miền tri thức *Sức khỏe*:

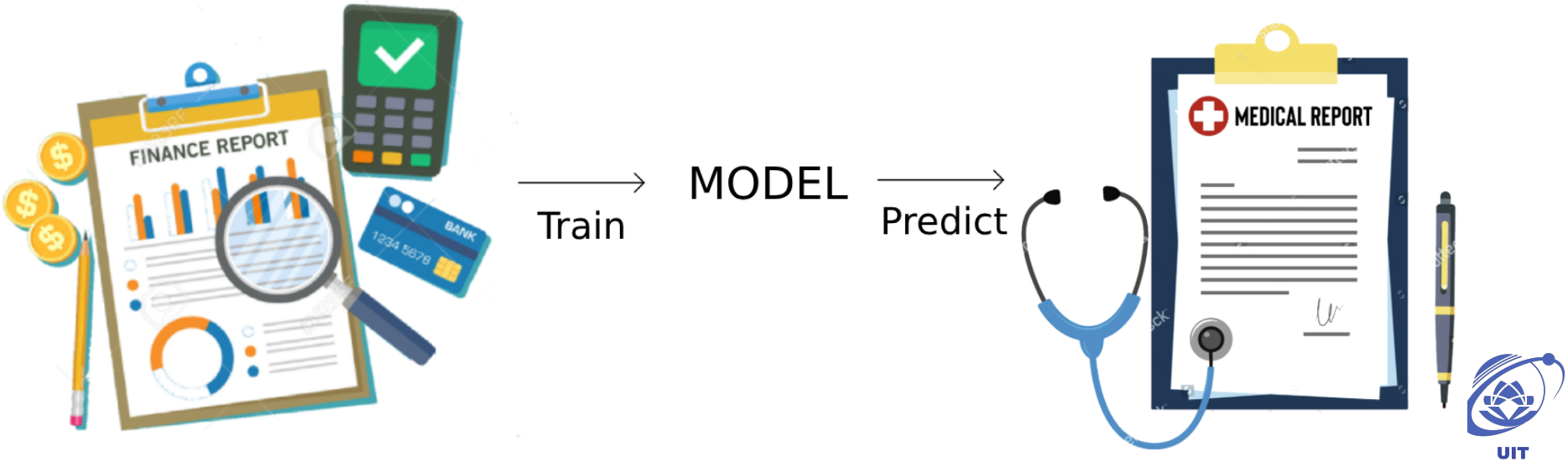
- Người bệnh ung thư tuyến giáp có các thể xuất hiện các dấu hiệu như khối u dai dẳng ở cổ, khó nuốt, thay đổi giọng nói.
- Phát hiện sớm thông qua các xét nghiệm chẩn đoán, nhất là người có nguy cơ di truyền bệnh, có thể cải thiện kết quả điều trị.

Nguồn: VnExpress

Bài toán tổng quát hóa miền tri thức

Định nghĩa

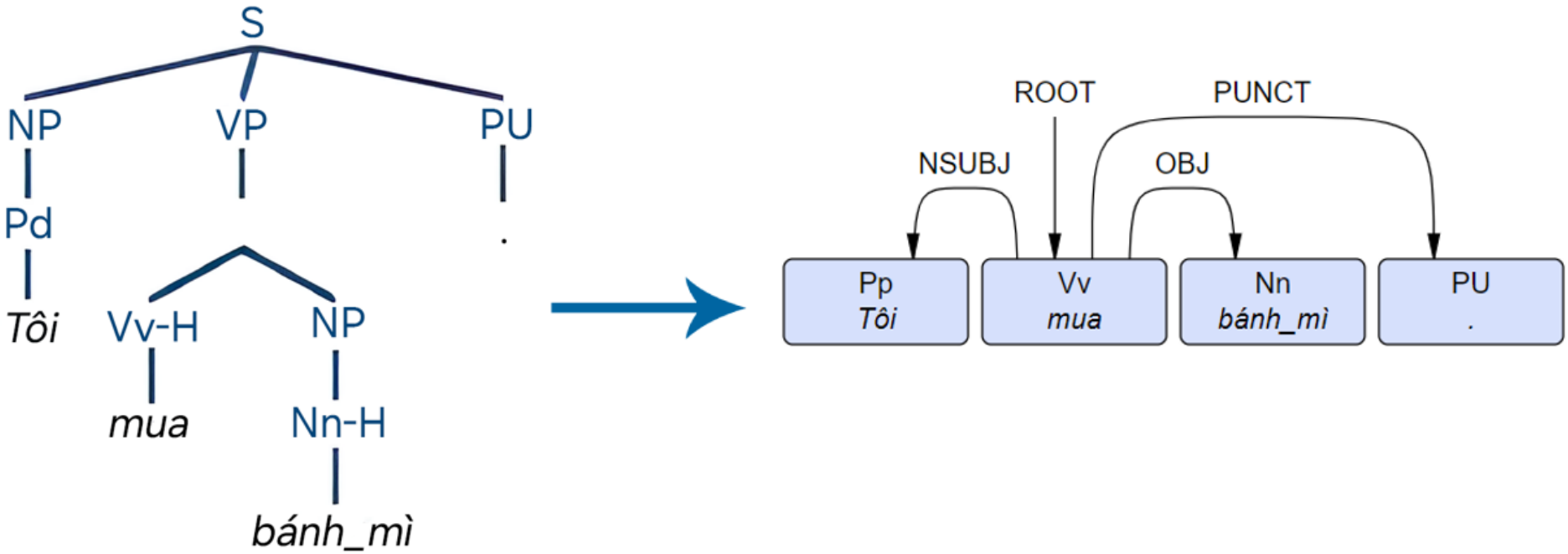
Tổng quát hóa miền tri thức (domain generalization) là một bài toán quan trọng trong học máy, mục tiêu là phát triển mô hình có khả năng áp dụng tri thức đã học được trong quá trình huấn luyện để xử lý hiệu quả tri thức mới chưa từng gặp mà không yêu cầu mô hình tiếp cận dữ liệu từ miền dữ liệu đích (target domain) trong quá trình huấn luyện.



Xây dựng kho ngữ liệu DGDT

Xây dựng kho ngữ liệu DGDT

Nhằm cân bằng giữa chất lượng và chi phí xây dựng kho ngữ liệu, chúng tôi áp dụng phương pháp chuyển đổi từ kho ngữ liệu cú pháp thành tố.



Lựa chọn kho ngữ liệu nguồn

Kho ngữ liệu nguồn: kho ngữ liệu NIIVTB-2 [1]

- Nguồn ngữ liệu có độ tin cậy cao, trích từ báo Thanh Niên.

[1] Nguyen, Q.T., Miyao, Y., Le, H.T.T. et al. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resources & Evaluation **52**, 269–315 (2018).

Lựa chọn kho ngữ liệu nguồn

Kho ngữ liệu nguồn: kho ngữ liệu NIIVTB-2 [1]

- Nguồn ngữ liệu có độ tin cậy cao, trích từ báo Thanh Niên.
- Bao gồm nhiều chủ đề khác nhau.

[1] Nguyen, Q.T., Miyao, Y., Le, H.T.T. et al. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resources & Evaluation **52**, 269–315 (2018).

Lựa chọn kho ngữ liệu nguồn

Kho ngữ liệu nguồn: kho ngữ liệu NIIVTB-2 [1]

- Nguồn ngữ liệu có độ tin cậy cao, trích từ báo Thanh Niên.
- Bao gồm nhiều chủ đề khác nhau.
- Mỗi chủ đề đóng vai trò là một miền tri thức.

Các chủ đề trong NIIVTB-2 được trích từ báo Thanh Niên

[1] Nguyen, Q.T., Miyao, Y., Le, H.T.T. et al. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resources & Evaluation **52**, 269–315 (2018).

Lựa chọn công cụ chuyển đổi tự động

Công cụ chuyển đổi tự động: công cụ của *Trương và cộng sự* [2] (converter).

- Tương thích với kho ngữ liệu cú pháp thành tố đã chọn.

[2] C. M. Truong, T. V. Pham, M. N. Phan, N. D. T. Le, T. V. Nguyen and Q. T. Nguyen, "Converting a constituency treebank to dependency treebank for Vietnamese,"2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 2022, pp. 256-261

Lựa chọn công cụ chuyển đổi tự động

Công cụ chuyển đổi tự động: công cụ của *Trương và cộng sự* [2] (converter).

- Tương thích với kho ngữ liệu cú pháp thành tố đã chọn.
- Đi kèm quy tắc dựng cây và bộ nhãn phụ thuộc phù hợp đặc trưng ngôn ngữ của tiếng Việt.

VnDT		**converter**
		ASUBJ
SUB	$\Rightarrow$	NSUBJ
		VSUBJ
		CSUBJ

So sánh giữa bộ nhãn của kho ngữ liệu VnDT và converter ở mối quan hệ danh từ.

[2] C. M. Truong, T. V. Pham, M. N. Phan, N. D. T. Le, T. V. Nguyen and Q. T. Nguyen, "Converting a constituency treebank to dependency treebank for Vietnamese,"2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 2022, pp. 256-261

Đánh giá chất lượng công cụ chuyển đổi

Để đảm bảo chất lượng của kho ngữ liệu, hai miền tri thức *Law* và *Life_of_youth* được sửa lỗi thủ công và so sánh với đầu ra của converter.

Miền tri thức	**Số lượng câu**	**UAS**	**LAS**
Law	610	95.55%	89.80%
Life_of_youth	635	95.69%	89.11%
Trung bình		95.62%	89.42%

Kết quả cho thấy đầu ra của converter có độ chính xác đủ cao để đảm bảo chất lượng ngữ liệu.

Thống kê của kho ngữ liệu DGDT

Tập	Miền tri thức	Số lượng câu	Số lượng từ
Train	Education	844	21.068
	Health	725	16.045
	Law	610	17.188
	Life_of_youth	635	16.396
	Military	690	16.949
	Politics_Society	712	20.245
	Science	692	16.964
	Sports	697	16.780
	Travel	540	13.397
	World	645	16.393
	Toàn tập	6.790	171.425
Dev	Entertainment	708	17.463
	Information_Technology	714	18.374
	Toàn tập	1.422	35.837
Test	Economic	725	18.207
	Life	828	19.537
	Toàn tập	1.553	37.744
Toàn kho ngữ liệu		9.765	245.006

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

So sánh với các kho ngữ liệu khác

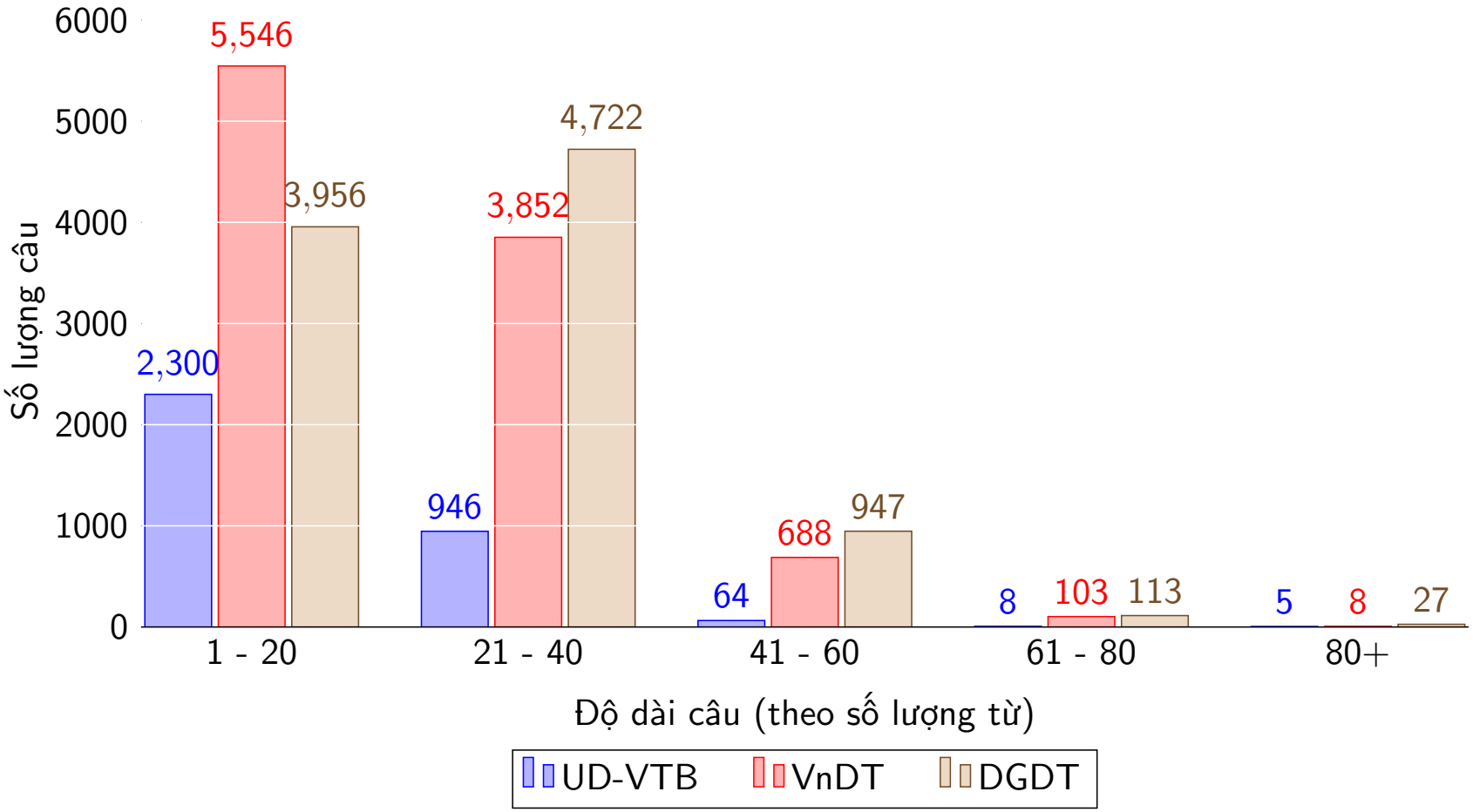
Kho ngữ liệu	Số lượng nhãn	Số lượng miền tri thức	Số lượng câu	Số lượng từ	Gán nhãn thủ công
BKTreebank [3]	26	không rõ	6,909	không rõ	✓
UD_Vietnamese-VTB [4]	84	1	3,323	58,069	✗
VnDT [5]	33	1	10,197	218,749	✗
DGDT	40	14	9,765	245,006	✗

[3] Nguyen, Hieu. (2017). BKTreebank: Building a Vietnamese Dependency Treebank. 10.48550/arXiv.1710.05519.

[4] https://github.com/UniversalDependencies/UD_Vietnamese-VTB

[5] https://github.com/datquocnguyen/VnDT

So sánh với các kho ngữ liệu khác



Xây dựng tiêu chuẩn đánh giá DGDTMark

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Tiêu chuẩn đánh giá DGDTMark

DGDT gồm 4 bối cảnh bối trí ngữ liệu nhằm đánh giá tác động của đa miền tri thức lên bài toán phân tích cú pháp phụ thuộc tiếng Việt.

1. **in-domain**: chia mỗi miền tri thức thành ba phần theo tỷ lệ 8:1:1 và gộp các phần tương ứng thành ba tập train, dev, test.

Tiêu chuẩn đánh giá DGDTMark

DGDT gồm 4 bối cảnh bối trí ngữ liệu nhằm đánh giá tác động của đa miền tri thức lên bài toán phân tích cú pháp phụ thuộc tiếng Việt.

1. **in-domain**: chia mỗi miền tri thức thành ba phần theo tỷ lệ 8:1:1 và gộp các phần tương ứng thành ba tập train, dev, test.
2. **domain-k-fold**: mỗi miền tri thức sẽ đóng vai trò làm tập dev và test, phần còn lại của kho ngữ liệu sẽ làm tập train.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Tiêu chuẩn đánh giá DGDTMark

DGDT gồm 4 bối cảnh bối trí ngữ liệu nhằm đánh giá tác động của đa miền tri thức lên bài toán phân tích cú pháp phụ thuộc tiếng Việt.

1. **in-domain**: chia mỗi miền tri thức thành ba phần theo tỷ lệ 8:1:1 và gộp các phần tương ứng thành ba tập train, dev, test.
2. **domain-k-fold**: mỗi miền tri thức sẽ đóng vai trò làm tập dev và test, phần còn lại của kho ngữ liệu sẽ làm tập train.
3. **domain-generalization**: bố trí ngữ liệu sao cho mỗi miền tri thức nằm độc nhất ở một trong ba tập train, dev, test.

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Tiêu chuẩn đánh giá DGDTMark

DGDT gồm 4 bối cảnh bối trí ngữ liệu nhằm đánh giá tác động của đa miền tri thức lên bài toán phân tích cú pháp phụ thuộc tiếng Việt.

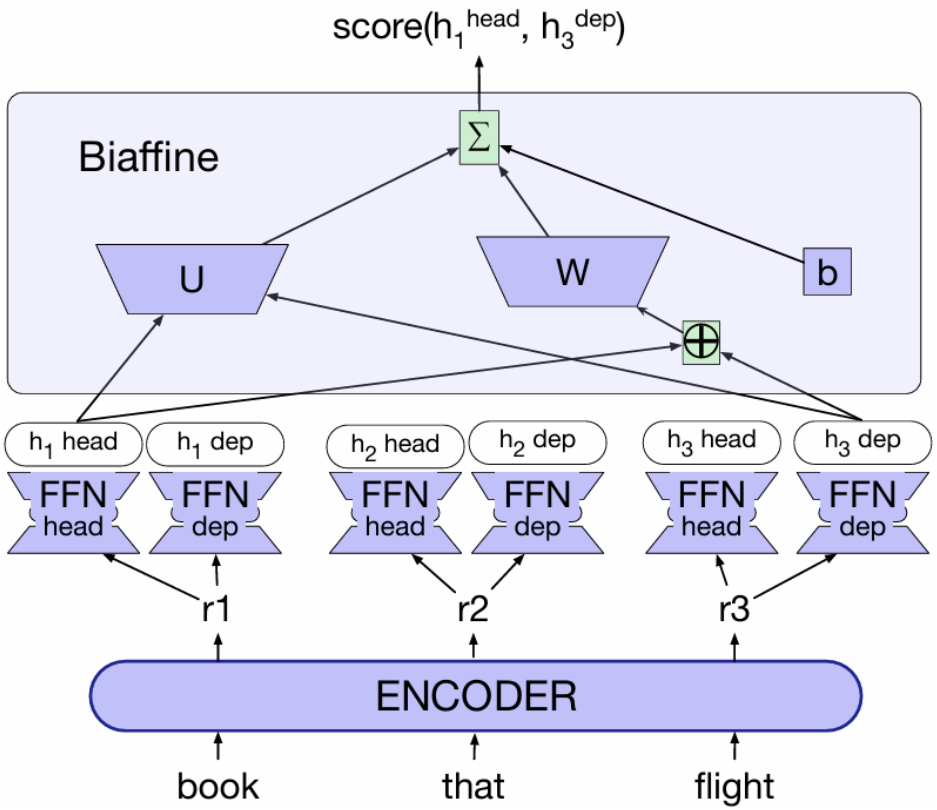
1. **in-domain**: chia mỗi miền tri thức thành ba phần theo tỷ lệ 8:1:1 và gộp các phần tương ứng thành ba tập train, dev, test.
2. **domain-k-fold**: mỗi miền tri thức sẽ đóng vai trò làm tập dev và test, phần còn lại của kho ngữ liệu sẽ làm tập train.
3. **domain-generalization**: bố trí ngữ liệu sao cho mỗi miền tri thức nằm độc nhất ở một trong ba tập train, dev, test.
4. **dataset-generalization**: bố trí ngữ liệu giống bối cảnh 3, thay tập test bằng tập test lấy từ kho ngữ liệu **NIIVTB_DT-1**.

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Mô hình xử lý bài toán

Baseline: Mô hình Deep Biaffine Parsing

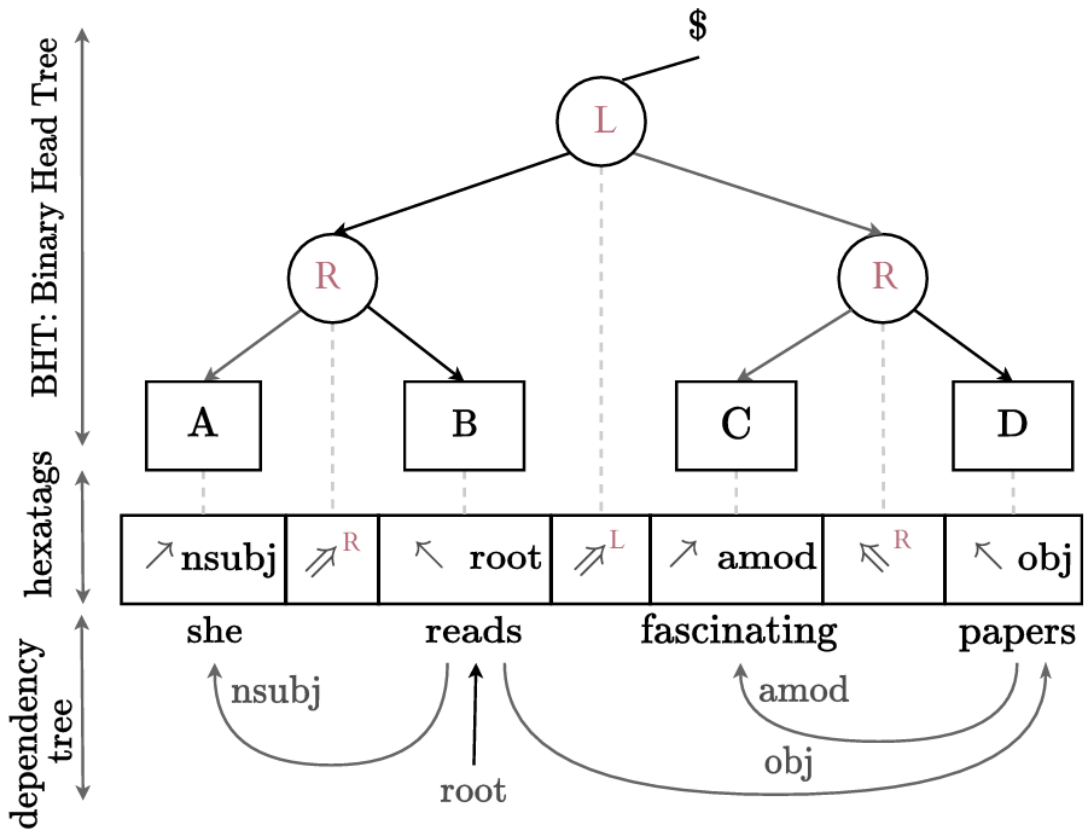
- Deep Biaffine Parsing (Biaffine) là mô hình có sức ảnh hưởng lớn đến sự phát triển của bài toán.
- Mô hình dựa trên hướng tiếp cận graph-based, trong đó sử dụng *biaffine attention* để tính điểm cho các cung trong đồ thị.



Kiến trúc của mô hình Biaffine

Baseline: Mô hình Hexatagger

- Trong mô hình này, mỗi cây phụ thuộc sẽ được chuyển đổi thành cây nhị phân (Binary Head Tree - BHT).
- Mô hình sẽ tiến hành học cách xây dựng BHT thông qua việc gán nhãn cho từng đỉnh trong cây.
- Sau khi thực hiện, mô hình giải mã BHT đã xây dựng lại thành cây phụ thuộc.

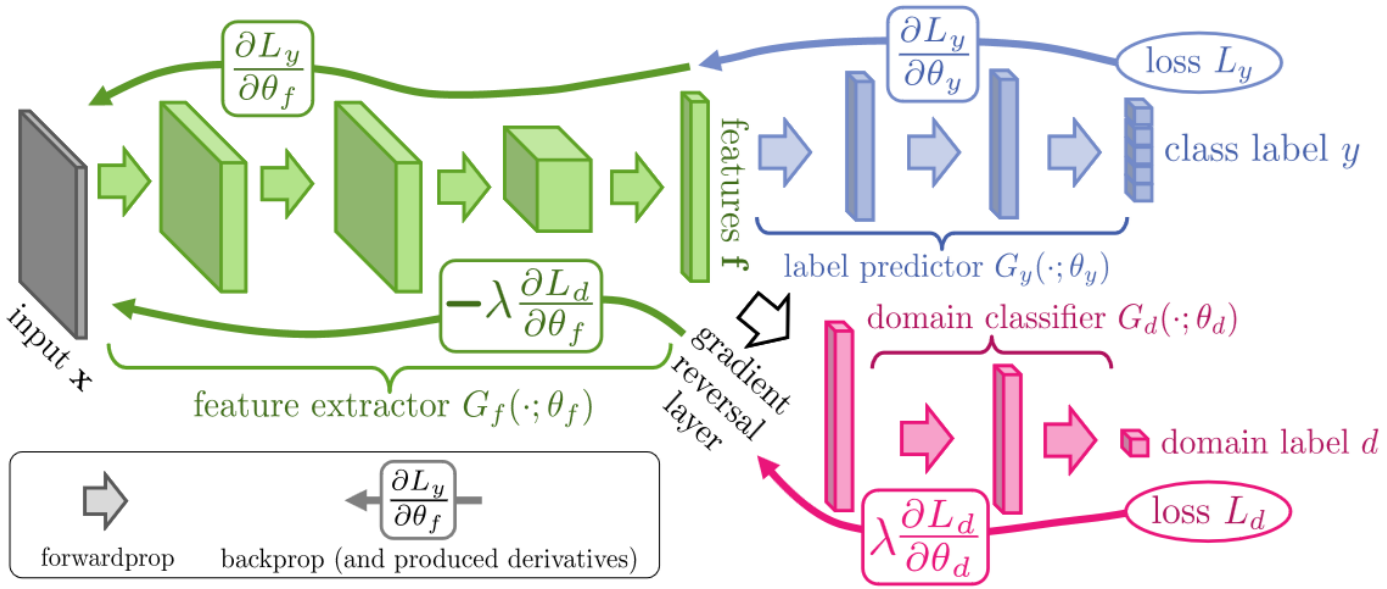


Phương pháp đề xuất: Adv

- Huấn luyện đối kháng *(Adversarial Training)* là một phương pháp được đề xuất nhằm giúp mô hình nắm bắt được các đặc trưng bất biến trên đa miền tri thức.

Phương pháp đề xuất: Adv

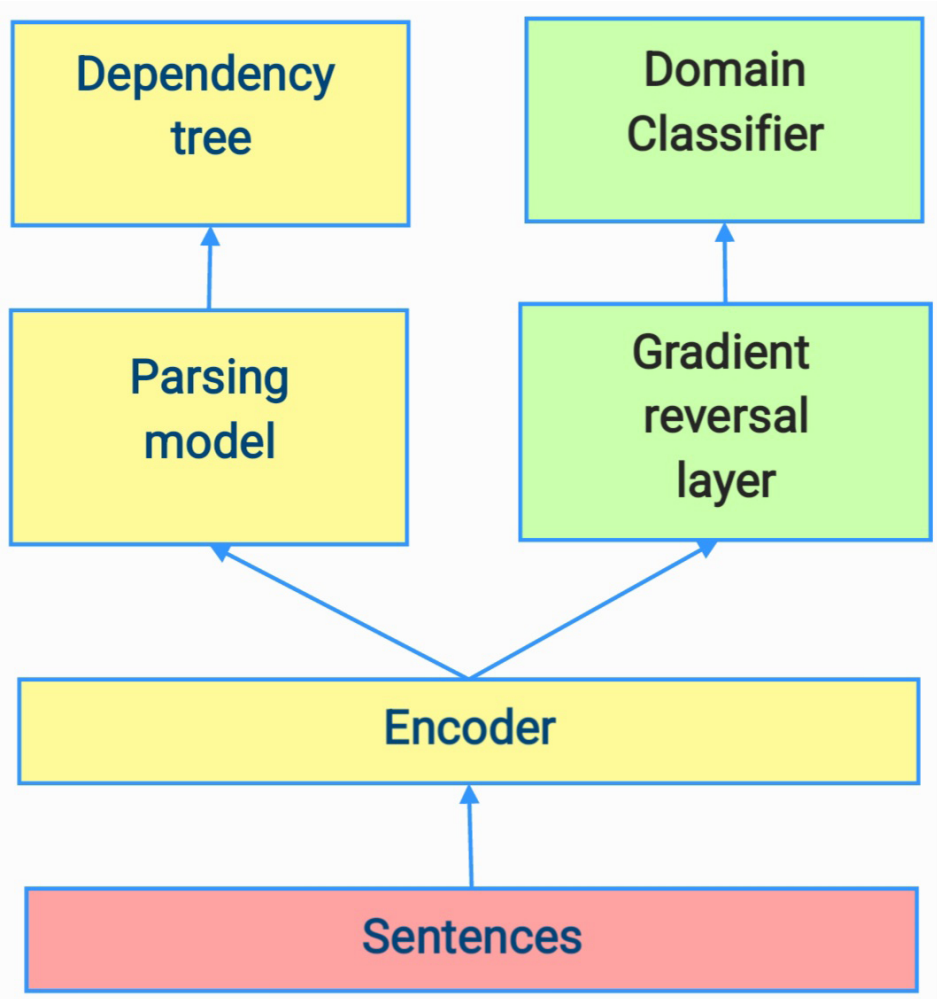
- Huấn luyện đối kháng *(Adversarial Training)* là một phương pháp được đề xuất nhằm giúp mô hình nắm bắt được các đặc trưng bất biến trên đa miền tri thức.
- Mô hình sẽ được huấn luyện song song với một bài toán khác (bài toán đối kháng) với giả thiết là khả năng mô hình xử lý tốt bài toán đối kháng **tỷ lệ nghịch** với khả năng mô hình nắm bắt được các đặt trưng bất biến.



Kiến trúc của phương pháp sử dụng huấn luyện đối kháng

Phương pháp đề xuất: Adv

- Bài đoán đối kháng được chọn là phân loại miền tri thức *(domain classifier)*.
- Lớp Gradient Reversal giúp **đảo ngược dấu** của đạo hàm trả về trong quá trình backward.
- Tham số $\gamma \in [0, 1]$ được thêm vào nhằm cân bằng giữa hai bài toán: $\mathcal{L}_{\text{model}} = \gamma \mathcal{L}_{\text{classifier}} + (1 - \gamma)\mathcal{L}_{\text{parser}}$



Kiến trúc mô hình do nhóm đề xuất

Thực nghiệm

Thiết lập thực nghiệm

- Baseline cho DGDTMark được xác lập thông qua hai mô hình là Biaffine và Hexatagger, mỗi mô hình được thực nghiệm trên hai encoder là PhoBERT và XLM-R.

Thiết lập thực nghiệm

- Baseline cho DGDTMark được xác lập thông qua hai mô hình là Biaffine và Hexatagger, mỗi mô hình được thực nghiệm trên hai encoder là PhoBERT và XLM-R.
- Phương pháp đề xuất được thực nghiệm trên mô hình Biaffine với encoder là PhoBERT.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Thiết lập thực nghiệm

- Baseline cho DGDTMark được xác lập thông qua hai mô hình là Biaffine và Hexatagger, mỗi mô hình được thực nghiệm trên hai encoder là PhoBERT và XLM-R.
- Phương pháp đề xuất được thực nghiệm trên mô hình Biaffine với encoder là PhoBERT.
- Mỗi mô hình được chạy trên 100 epoch, mô hình tốt nhất được lựa chọn trên tập dev với độ đo LAS.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Kết quả thực nghiệm trên DGDTMark

Độ đo: **UAS (%)**

Mô hình	Bộ mã hoá	*in-domain*	*domain-k-fold*	*domain-generalization*	*dataset-generalization*
Biaffine	PhoBERT	**91.62**	**91.22**	88.95	88.54
	XLM-R	90.25	89.72	88.26	86.98
Hexatagger	PhoBERT	91.49	91.00	**90.00**	**89.19**
	XLM-R	89.72	89.97	89.62	89.01

Độ đo: **LAS (%)**

Mô hình	Bộ mã hoá	*in-domain*	*domain-k-fold*	*domain-generalization*	*dataset-generalization*
Biaffine	PhoBERT	87.83	87.01	84.76	82.74
	XLM-R	85.89	84.99	82.88	80.97
Hexatagger	PhoBERT	**89.00**	**88.28**	**86.81**	**85.40**
	XLM-R	86.99	87.08	86.48	85.22

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Kết quả thực nghiệm trên *domain-k-fold* (UAS)

Miền tri thức	Biaffine	Hexatagger
Economic	**90.53**	90.21
Education	**90.97**	90.92
Entertainment	90.52	**90.57**
Health	**91.18**	91.07
Information_Technology	**90.36**	90.28
Law	**91.58**	91.05
Life	89.50	**89.91**
Life_of_youth	**91.12**	90.85
Military	**92.29**	92.16
Politics_Society	**91.16**	90.43
Science	**92.00**	91.71
Sports	**91.84**	91.49
Travel	**91.17**	90.67
World	**92.89**	92.62
Trung bình	**91.22**	91.00

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Kết quả thực nghiệm trên *domain-k-fold* (LAS)

Miền tri thức	Biaffine	Hexatagger
Economic	85.82	**87.32**
Education	86.74	**88.05**
Entertainment	85.94	**87.47**
Health	86.55	**87.99**
Information_Technology	85.89	**87.43**
Law	87.45	**88.49**
Life	84.82	**86.83**
Life_of_youth	86.56	**87.82**
Military	88.82	**89.89**
Politics_Society	86.83	**87.69**
Science	88.85	**89.63**
Sports	87.16	**88.69**
Travel	87.03	**87.97**
World	89.61	**90.60**
Trung bình	87.01	**88.28**

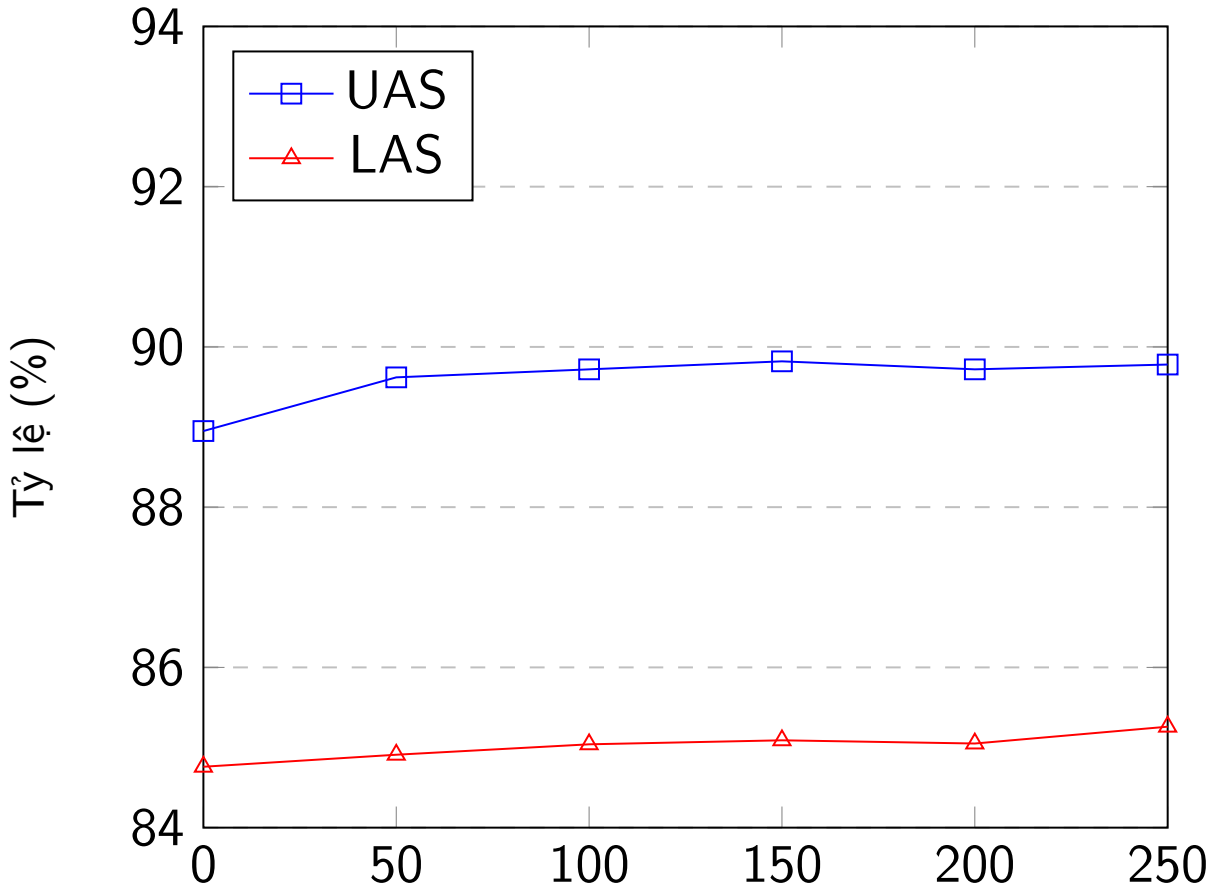
UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Phân tích hiệu suất theo nhãn

Nhãn phụ thuộc	Số lần xuất hiện	UAS (%)	LAS (%)
PUNCT	5136	83.94	83.92
NN	3570	92.83	88.01
OBJ	3080	94.42	90.06
VMOD	2637	91.13	81.53
PREP	2570	85.37	83.66
NSUBJ	2340	90.64	89.53
POBJ	2216	**97.70**	95.35
ADJUNCT	2185	93.00	90.30
CONJ	2169	78.01	74.09
ROOT	1553	90.86	90.86
CC	1250	82.56	81.84
AMOD	1132	89.66	85.42
DET	1062	96.23	**95.10**

Danh sách bao gồm những nhãn xuất hiện hơn 1000 lần trong tập đánh giá.

Ảnh hưởng của dữ liệu kiểm tra trong quá trình huấn luyện



Kết quả thực nghiệm phương pháp đề xuất trên DGDTMark

Độ đo: **UAS (%)**

Mô hình	*in-domain*	*domain-k-fold*	*domain-generalization*	*dataset-generalization*
Biaffine	**91.62**	**91.22**	88.95	**88.54**
Biaffine+Adv	91.20	90.89	**89.41**	87.51

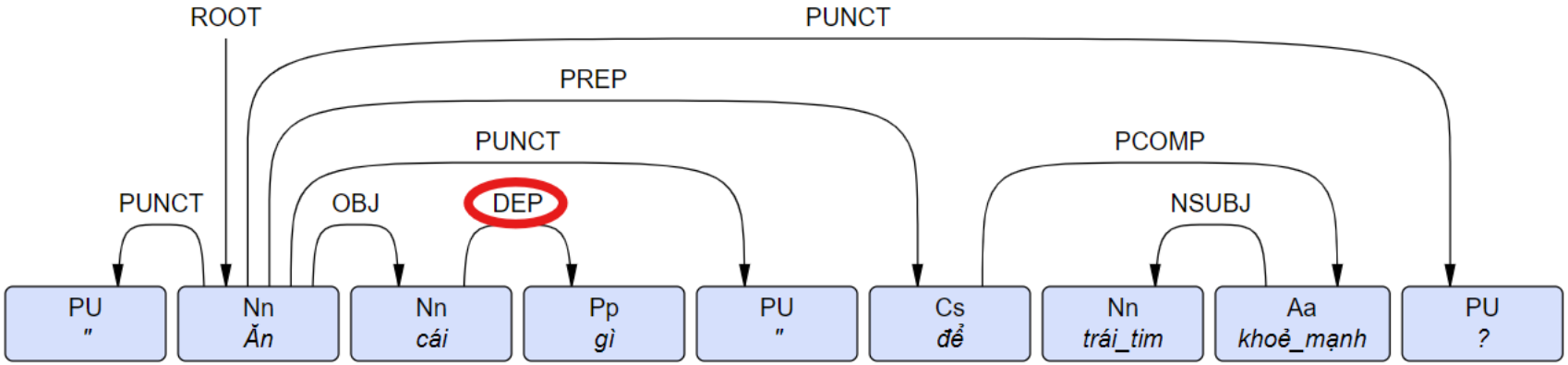
Độ đo: **LAS (%)**

Mô hình	*in-domain*	*domain-k-fold*	*domain-generalization*	*dataset-generalization*
Biaffine	87.83	87.01	84.76	**82.74**
Biaffine+Adv	**87.88**	**87.22**	**85.25**	82.62

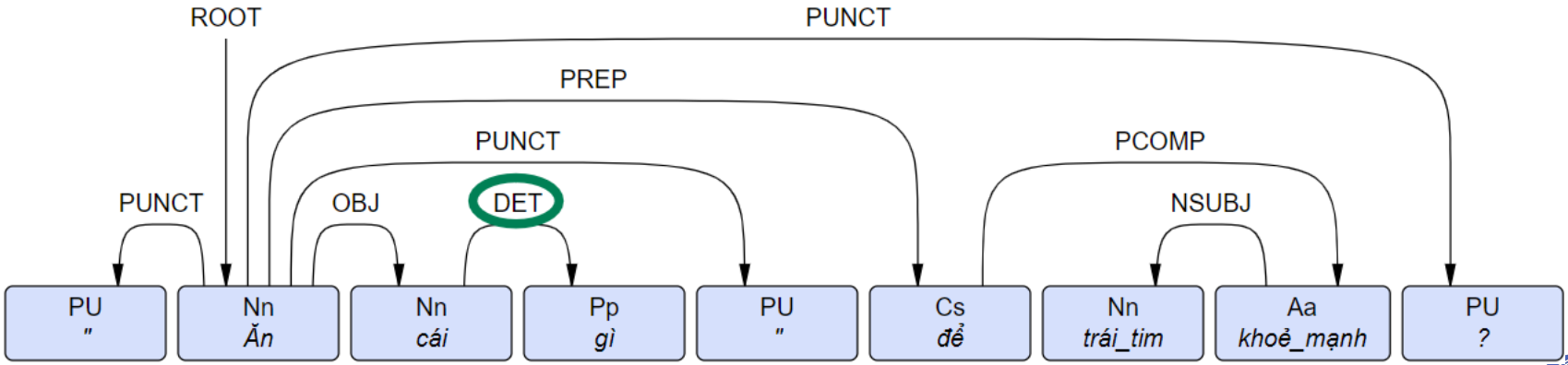
UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

So sánh dự đoán trên bối cảnh *in-domain*

- Dự đoán của mô hình Biaffine (sai):

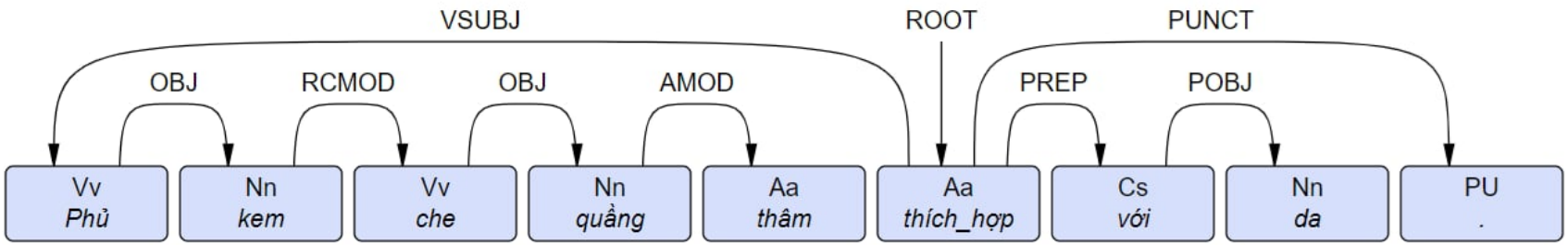


- Dự đoán của mô hình đề xuất Biafine+Adv (đúng):

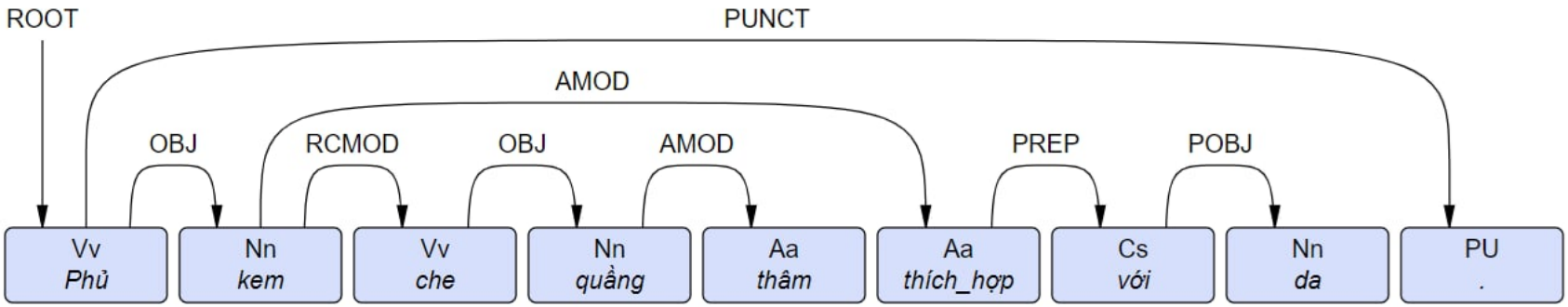


So sánh dự đoán trên bối cảnh *domain-generalization*

- Dự đoán của mô hình Biaffine (sai):



- Dự đoán của mô hình đề xuất Biafine+Adv (đúng):



Công bố khoa học

Công bố khoa học

[1] Vinh-Hien D. Huynh, Chau-Anh Le, Chau M. Truong, Y Thien Huynh and Quy T. Nguyen, *Domain Generalization in Vietnamese Dependency Parsing: A Novel Benchmark and Domain Gap Analysis*, In Proceedings of the 13th International Symposium on Information and Communication Technology (SOICT '24)

⇒ Trình bày về việc xây dựng kho ngữ liệu DGDT và tiêu chuẩn đánh giá DGDTMark, thiết lập baseline bằng mô hình Biaffine và phân tích ảnh hưởng của domain gap.

Kết luận

Kết luận

- Khoá luận đã xây dựng kho ngữ liệu DGDT, một kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt trên đa miền tri thức.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Kết luận

- Khoá luận đã xây dựng kho ngữ liệu DGDT, một kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt trên đa miền tri thức.
- Ngoài ra, nhóm nghiên cứu đã đề xuất tiêu chuẩn đánh giá DGDTMark nhằm đánh giá tác động của domain gap lên bài toán phân tích cú pháp phụ thuộc.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Kết luận

- Khoá luận đã xây dựng kho ngữ liệu DGDT, một kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt trên đa miền tri thức.
- Ngoài ra, nhóm nghiên cứu đã đề xuất tiêu chuẩn đánh giá DGDTMark nhằm đánh giá tác động của domain gap lên bài toán phân tích cú pháp phụ thuộc.
- Khoá luận cũng giới thiệu một hướng tiếp cận mới nhằm cải thiện khả năng tổng quát hoá miền tri thức của các mô hình.

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

Kết luận

- Khoá luận đã xây dựng kho ngữ liệu DGDT, một kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt trên đa miền tri thức.
- Ngoài ra, nhóm nghiên cứu đã đề xuất tiêu chuẩn đánh giá DGDTMark nhằm đánh giá tác động của domain gap lên bài toán phân tích cú pháp phụ thuộc.
- Khoá luận cũng giới thiệu một hướng tiếp cận mới nhằm cải thiện khả năng tổng quát hoá miền tri thức của các mô hình.

→ Kết quả thực nghiệm cho thấy, các mô hình tiên tiến nhất hiện nay (SOTA) vẫn chịu ảnh hưởng đáng kể từ vấn đề domain gap.

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Kết luận

- Khoá luận đã xây dựng kho ngữ liệu DGDT, một kho ngữ liệu phân tích cú pháp phụ thuộc tiếng Việt trên đa miền tri thức.
- Ngoài ra, nhóm nghiên cứu đã đề xuất tiêu chuẩn đánh giá DGDTMark nhằm đánh giá tác động của domain gap lên bài toán phân tích cú pháp phụ thuộc.
- Khoá luận cũng giới thiệu một hướng tiếp cận mới nhằm cải thiện khả năng tổng quát hoá miền tri thức của các mô hình.

→ Kết quả thực nghiệm cho thấy, các mô hình tiên tiến nhất hiện nay (SOTA) vẫn chịu ảnh hưởng đáng kể từ vấn đề domain gap.

→ Phương pháp đề xuất trong khoá luận đã thể hiện tiềm năng trong việc cải thiện hiệu suất của mô hình ở một số bối cảnh bố trí ngữ liệu.

UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Q&A

UIT
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN