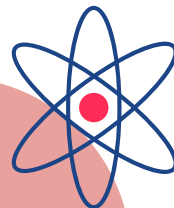




NGŨ NGHĨA HỌC TÍNH TOÁN



GÁN NHÃN NGHĨA CỦA TỪ



GVHD: Ph.D Nguyễn Trọng Chinh

Nhóm: VungTauCity

Đặng Thanh Ngân

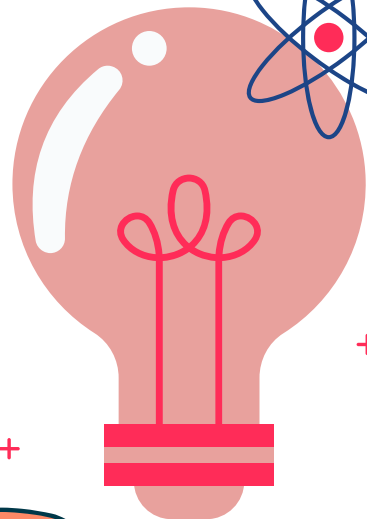
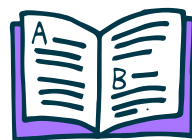
22520929

Hà Ngũ Long Nguyễn

22520965

Phạm Thanh Thảo

22521373





01

GIỚI THIỆU

Mục tiêu đề tài
Đoạn văn
Các câu truy vấn

04

THỰC NGHIỆM

Cài đặt Prolog
Xử lý câu truy vấn
Ứng dụng



02

NGÔN NGỮ BẬC MỘT

Các từ vựng
Mô hình $\langle D, F \rangle$
Dịch sang FOL

05

KẾT LUẬN & NHẬN XÉT

Nhận xét
Hạn chế
Kết luận

03

GÁN NHÃN NGHĨA CỦA TỪ

Tiền xử lý dữ liệu
Huấn luyện mô hình
Gán nhãn thủ công



01

GIỚI THIỆU

- Mục tiêu đề tài
- Giới thiệu Đoạn văn
- Các câu truy vấn





1. GIỚI THIỆU

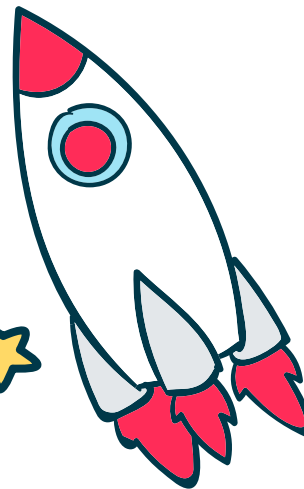
1.1. Mục tiêu đề tài

Bổ sung tri thức từ WordNet, bao gồm:

- Tra cứu nghĩa của từ theo danh sách synset.
- Xác định quan hệ ngữ nghĩa như hypernym (quan hệ bao), hyponym (quan hệ loại), meronym (quan hệ bộ phận).
- Bổ sung tri thức mới vào cơ sở tri thức hiện tại.

Truy vấn tri thức theo ngôn ngữ bậc một:

- Trả lời câu hỏi dạng xác nhận (Yes/No).
- Truy vấn đối tượng.
- Suy luận dựa trên tri thức bổ sung từ WordNet.



1.2. Giới thiệu Đoạn văn



Ngan, Nguyen, and Thao are students at UIT. They are friends. Nguyen is from Gia Lai, Ngan and Thao were born in the coastal city of Vung Tau. Gia Lai is lush. The beach in Vung Tau is beautiful and refreshing. Everyone loves this beach. Thao's dad is a fisherman, and Thao's mom is a housewife. Ngan's dad and Ngan's mom are fishmongers. Nguyen's dad and Nguyen's mom are farmers. They love their hometowns.

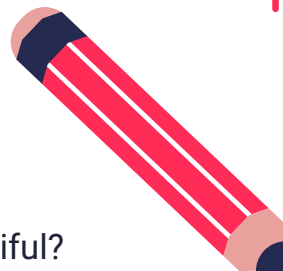




1. GIỚI THIỆU

1.2. Một số câu truy vấn

- Q1: Who is a student at UIT?
- Q2: Is Nguyen from Vung Tau?
- Q3: Are Ngan and Thao from the same coastal city?
- Q4: Is Thao's mom a homemaker?
- Q5: Does Nguyen's dad work with fish?
- Q6: Who is from Gia Lai?
- Q7: Which city are Ngan and Thao from?
- Q8: Is Vung Tau a coastal city?
- Q9: Who is a fisher?
- Q10: What is Nguyen's mom job?
- Q11: Is Ngan's mom a fishmonger?
- Q12: Do Ngan and Thao adore Vung Tau?
- Q13: Are Ngan and Thao friends with Nguyen?



- Q14: Is the beach in Vung Tau beautiful?
- Q15: Is the beach in Vung Tau fresh?
- Q16: Does everyone love Vung Tau's beach?
- Q17: Are Ngan and Thao school friends?
- Q18: Are Nguyen's dad farmer?
- Q19: Does Nguyen care for Vung Tau beach?
- Q20: Are Ngan and Thao pupil at Bach Khoa university?
- Q21: An is a person. Does An adore Vung Tau beach?
- Q22: Ngan is a student at UIT. Does she love her hometown?
- Q23: Vung Tau is a coastal city. Is beach in that beautiful and refreshing?
- Q24: Thao has a father and a mother. Is her father a fisherman and her mother a housewife?

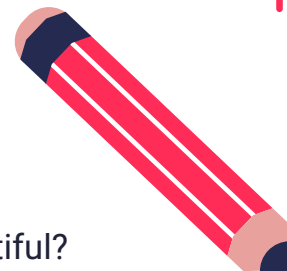




1. GIỚI THIỆU

1.2. Một số câu truy vấn

- Q1: Who is a student at UIT?
- Q2: Is Nguyen from Vung Tau?
- Q3: Are Ngan and Thao from the same coastal city?
- Q4: Is Thao's mom a homemaker?
- Q5: Does Nguyen's dad work with fish?
- Q6: Who is from Gia Lai?
- Q7: Which city are Ngan and Thao from?
- Q8: Is Vung Tau a coastal city?
- Q9: Who is a fisher?
- Q10: What is Nguyen's mom job?
- Q11: Is Ngan's mom a fishmonger?
- Q12: Do Ngan and Thao adore Vung Tau?
- Q13: Are Ngan and Thao friends with Nguyen?



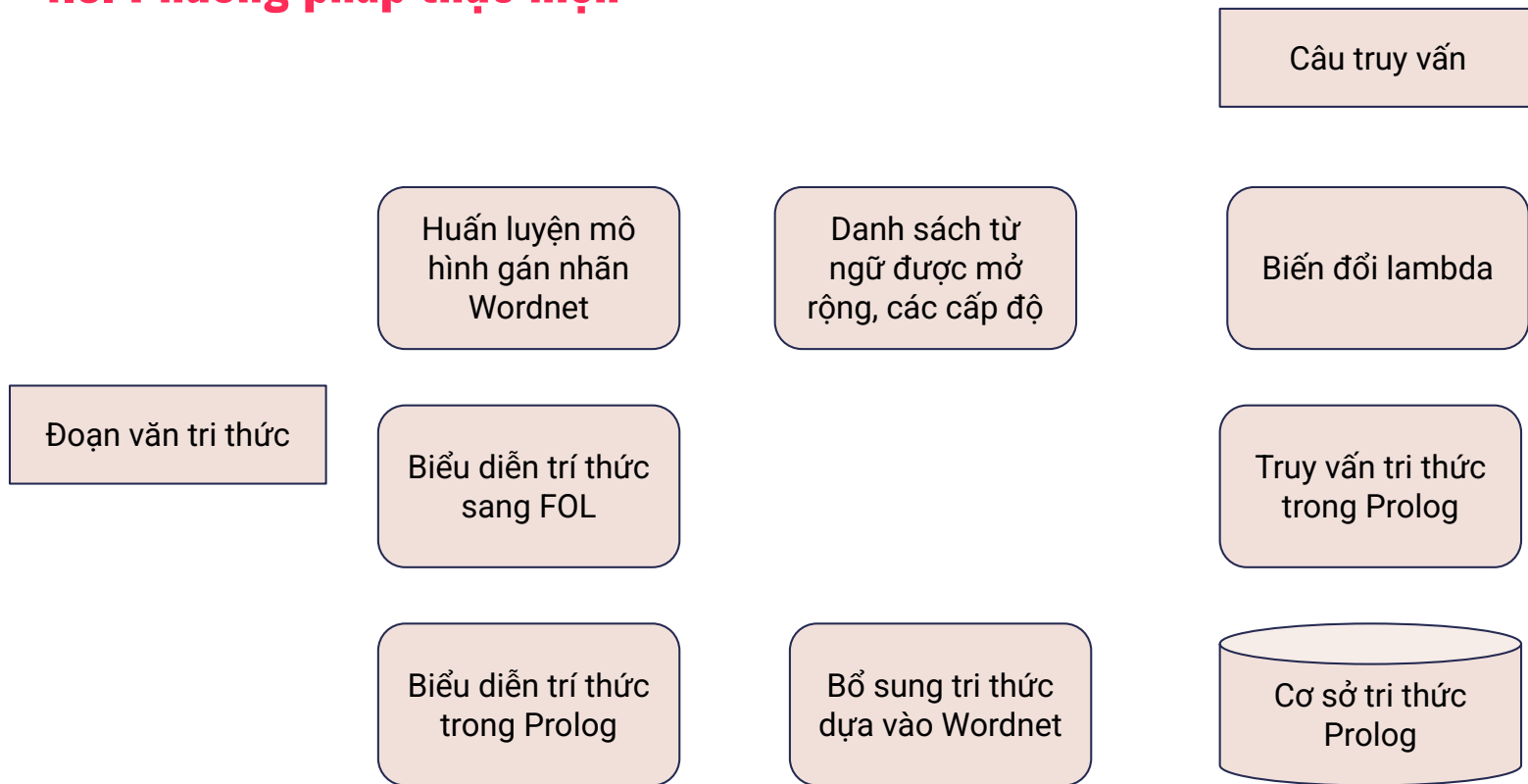
- Q14: Is the beach in Vung Tau beautiful?
- Q15: Is the beach in Vung Tau fresh?
- Q16: Does everyone love Vung Tau's beach?
- Q17: Are Ngan and Thao school friends?
- Q18: Are Nguyen's dad farmer?
- Q19: Does Nguyen care for Vung Tau beach?
- Q20: Are Ngan and Thao pupil at Bach Khoa university?
- Q21: An is a person. Does An adore Vung Tau beach?
- Q22: Ngan is a student at UIT. Does she love her hometown?
- Q23: Vung Tau is a coastal city. Is beach in that beautiful and refreshing?
- Q24: Thao has a father and a mother. Is her father a fisherman and her mother a housewife?

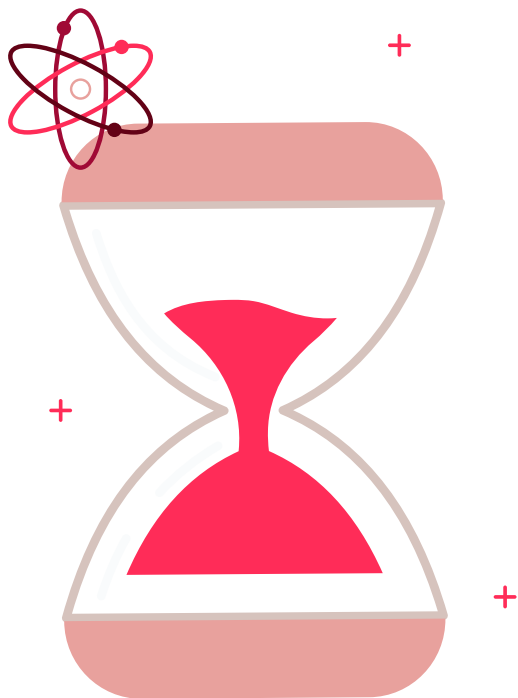




1. GIỚI THIỆU

1.3. Phương pháp thực hiện





02

NGÔN NGỮ BẬC MỘT

Các từ vựng
Mô hình $\langle D, F \rangle$
Dịch sang FOL



2. NGÔN NGỮ BẬC MỘT



2.1. Từ vựng bậc một

Ngan, Nguyen, and Thao are students at UIT. They are friends. Nguyen is from Gia Lai, Ngan and Thao were born in the coastal city of Vung Tau. Gia Lai is lush. The beach in Vung Tau is beautiful and refreshing. Everyone loves this beach. Thao's dad is a fisherman, and Thao's mom is a housewife. Ngan's dad and Ngan's mom are fishmongers. Nguyen's dad and Nguyen's mom are farmers. They love their hometowns.



$V = \{\text{Ngan}(0), \text{Thao}(0), \text{Nguyen}(0), \text{UIT}(0), \text{GiaLai}(0), \text{VungTau}(0), \text{coastal}(1), \text{lush}(1), \text{beach}(1), \text{refreshing}(1), \text{lush}(1), \text{person}(1), \text{student}(1), \text{beautiful}(1), \text{dad}(1), \text{mom}(1), \text{fisherman}(1), \text{housewife}(1), \text{fishmonger}(1), \text{farmer}(1), \text{at}(2), \text{friend}(2), \text{from}(2), \text{be_born}(2), \text{in}(2), \text{love}(2)\}$



2. NGÔN NGỮ BẬC MỘT

2.1. Mô hình <D,F>

Ngan, Nguyen, and Thao are students at UIT. They are friends. Nguyen is from Gia Lai, Ngan and Thao were born in the coastal city of Vung Tau. Gia Lai is lush. The beach in Vung Tau is beautiful and refreshing. Everyone loves this beach. Thao's dad is a fisherman, and Thao's mom is a housewife. Ngan's dad and Ngan's mom are fishmongers. Nguyen's dad and Nguyen's mom are farmers. They love their hometowns.

Miền đối tượng (D):

$D = \{\text{Ngan, Nguyen, Thao, UIT, GL, VT, BH, TD, TM, NM, ND, LD, LM}\}$

Tập hợp Ký hiệu Hàm (F_{logic}):

$F = \{$

$f(\text{student}) = \{\text{Ngan, Nguyen, Thao}\}, f(\text{at}) = \{(\text{Ngan, UIT}), (\text{Nguyen, UIT}), (\text{Thao, UIT})\},$

$f(\text{friend}) = \{(\text{Ngan, Nguyen}), (\text{Nguyen, Ngan}), (\text{Nguyen, Thao}), (\text{Thao, Nguyen}), (\text{Ngan, Thao}), (\text{Thao, Ngan})\},$

$f(\text{from}) = \{(\text{Nguyen, GL})\}, f(\text{be_born}) = \{(\text{Ngan, VT}), (\text{Thao, VT})\},$

$f(\text{dad}) = \{(\text{TD}), (\text{ND}), (\text{LD})\}, f(\text{mom}) = \{(\text{TM}), (\text{NM}), (\text{LM})\},$

$f(s) = \{(\text{Thao, TD}), (\text{Thao, TM}), (\text{Nguyen, LD}), (\text{Nguyen, LM}),$

$(\text{Ngan, ND}), (\text{Ngan, NM})\},$

$f(\text{fisherman}) = \{(\text{TD})\}$

$f(\text{housewife}) = \{(\text{TM})\}$

$f(\text{fishmonger}) = \{(\text{NM}), (\text{ND})\}$

$f(\text{farmer}) = \{(\text{LM}), (\text{LD})\},$

$f(\text{lush}) = \{(\text{GL})\}$

$f(\text{hometown}) = \{(\text{Nguyen, GL}), (\text{Ngan, VT}), (\text{Thao, VT})\},$

$f(\text{coastal}) = \{\text{VT}\}, f(\text{city}) = \{\text{VT, GL}\},$

$f(\text{beach}) = \{\text{BH}\}, f(\text{in}) = \{(\text{BH, VT})\}$

$f(\text{beautiful}) = \{\text{BH}\}, f(\text{refreshing}) = \{\text{BH}\},$

$f(\text{love}) = \{(\text{Thao, VT}), (\text{Ngan, VT}),$

$(\text{Nguyen, GL})\}$



1. GIỚI THIỆU

1.3. Dịch sang FOL



Câu 1: Ngan, Nguyen, and Thao are students at UIT.

Logic: $F_1 = (\text{student}(\text{Ngan}) \wedge \text{at}(\text{Ngan}, \text{UIT})) \wedge (\text{student}(\text{Thao}) \wedge \text{at}(\text{Thao}, \text{UIT})) \wedge (\text{student}(\text{Nguyen}) \wedge \text{at}(\text{Nguyen}, \text{UIT}))$

Câu 2: They are friends. (Giả định mối quan hệ bạn bè là đối xứng)

Logic: $F_2 = (\text{friend}(\text{Ngan}, \text{Thao}) \wedge \text{friend}(\text{Thao}, \text{Ngan})) \wedge (\text{friend}(\text{Ngan}, \text{Nguyen}) \wedge \text{friend}(\text{Nguyen}, \text{Ngan})) \wedge (\text{friend}(\text{Thao}, \text{Nguyen}) \wedge \text{friend}(\text{Nguyen}, \text{Thao}))$

Câu 3: But Nguyen is from Gia Lai, Ngan and Thao were born in the coastal city of Vung Tau.

Logic: $F_3 = \text{from}(\text{Nguyen}, \text{GL}) \wedge \text{city}(\text{GL}) \wedge \text{be_born}(\text{Ngan}, \text{VT}) \wedge \text{be_born}(\text{Thao}, \text{VT}) \wedge \text{coastal_city}(\text{VT}) \wedge \text{city}(\text{VT})$

Câu 4: Gia Lai is lush.

Logic: $F_4 = \text{lush}(\text{GL})$

Câu 5: The beach in Vung Tau is beautiful and refreshing.

Logic: $F_5 = \text{beach}(\text{BH}) \wedge \text{beautiful}(\text{BH}) \wedge \text{refreshing}(\text{BH})$

Câu 6: Everyone loves Vung Tau's beaches. (Giả định 'Everyone' là 'person(y)')

Logic: $F_6 = \forall y, \text{person}(y) \wedge \text{beach}(\text{BH}) \wedge \text{in}(\text{BH}, \text{Vung_Tau}) \rightarrow \text{loves}(y, \text{BH})$





1. GIỚI THIỆU

1.1. Mục tiêu đề tài



Câu 7: Thao's dad is a fisherman, and Thao's mom is a housewife.

Logic: $F_7 = \text{dad}(\text{TD}) \wedge \text{s}(\text{Thao}, \text{TD}) \wedge \text{fisherman}(\text{TD}) \wedge \text{mom}(\text{TM}) \wedge \text{s}(\text{Thao}, \text{TM}) \wedge \text{housewife}(\text{TM})$

Câu 8: Ngan's dad and Ngan's mom are fishmonger. (Giả định mối quan hệ bạn bè là đối xứng)

Logic: $F_8 = \text{dad}(\text{ND}) \wedge \text{s}(\text{Ngan}, \text{ND}) \wedge \text{fishmonger}(\text{ND}) \wedge \text{mom}(\text{NM}) \wedge \text{s}(\text{Ngan}, \text{NM}) \wedge \text{fishmonger}(\text{NM})$

Câu 9: Nguyen's dad and Nguyen's are farmers.

Logic: $F_9 = \text{dad}(\text{LD}) \wedge \text{s}(\text{Nguyen}, \text{LD}) \wedge \text{farmer}(\text{LD}) \wedge \text{mom}(\text{LM}) \wedge \text{s}(\text{Nguyen}, \text{LM}) \wedge \text{farmer}(\text{LM})$

Câu 10: They love their hometown.

Logic: $F_{10} = \text{love}(\text{Thao}, \text{VT}) \wedge \text{love}(\text{Ngan}, \text{VT}) \wedge \text{love}(\text{Nguyen}, \text{GL})$



Cơ sở tri thức (KB) được tạo thành từ phép Hội (\wedge) của tất cả các công thức logic từ F1 đến F10:

$\text{KB} = F_1 \wedge F_2 \wedge F_3 \wedge F_4 \wedge F_5 \wedge F_6 \wedge F_7 \wedge F_8 \wedge F_9 \wedge F_{10}$

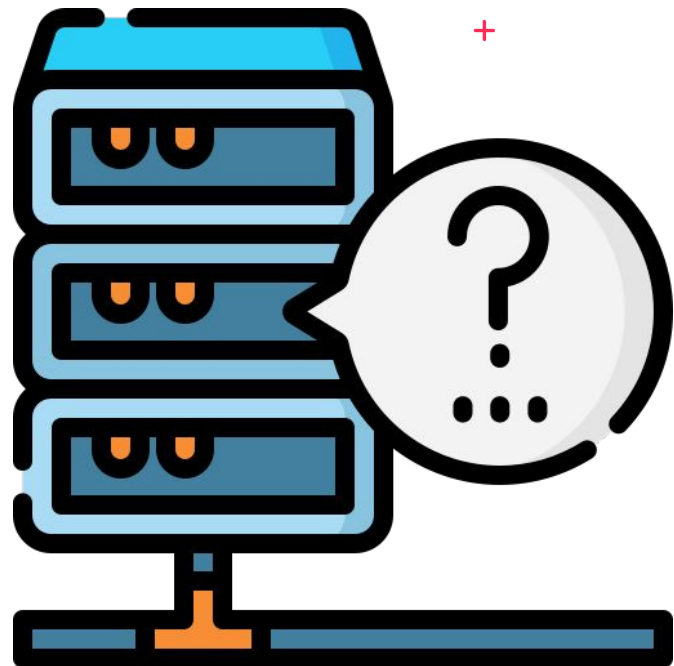


03

GÁN NHÃN

NGHĨA CỦA TỪ

- Giới thiệu bài toán
- Gán nhãn thủ công
- Tiền xử lý dữ liệu
- Huấn luyện mô hình



+

3. Gán nhãn nghĩa của từ

3.1. Giới thiệu bài toán

- **Input:** Một câu văn bản tự nhiên.
- **Output:** Câu văn bản tương ứng, trong đó các từ có nghĩa quan trọng được gán thêm nhãn sense key WordNet.

Input	Output
the	X
beach	beach%1:17:00::
in	X
Vung_tau	location%1:03:00::
is	be%2:42:03::
beautiful	beautiful%3:00:00::
and	X
refreshing	refreshing%5:00:00::

beach/**beach%1:17:00::**

- Từ: beach
- POS: 1 → danh từ (noun)
- Định nghĩa (definition): “an area of sand sloping down to the water of a sea or lake.” → dịch: “một vùng cát thoải dốc xuống mặt nước của biển hoặc hồ.”
- Ví dụ trong WordNet: “the children played all day on the beach.”
- WordNet synset: beach.n.01
- WordNet file mapping: %1:17:00::
 - → 1: noun
 - → 17: noun.object
 - → 00: sense id 0 (nghĩa đầu tiên)

+ *X: thể hiện từ không có trong wordnet

+



3. Gán nhãn nghĩa của từ



3.1. Giới thiệu bài toán

Ý nghĩa của việc gán nhãn WordNet:

- Tạo dữ liệu chuẩn (gold standard) phục vụ huấn luyện và đánh giá mô hình NLP.
- Giúp phân giải nghĩa từ chính xác trong ngữ cảnh, giảm mơ hồ đa nghĩa (ambiguous).
- Chuẩn hóa biểu diễn ngữ nghĩa theo WordNet synset, dễ so sánh và tái sử dụng.
- Nâng cao hiệu quả các bài toán NLP như dịch máy, tìm kiếm ngữ nghĩa, hỏi–đáp.
- Hỗ trợ phân tích ngôn ngữ học và nghiên cứu ngữ nghĩa.
- Tăng tính minh bạch và khả năng giải thích của hệ thống xử lý ngôn ngữ.

Ví dụ 1: từ bank (đa nghĩa)

He sat on the bank and watched the river.

- bank có nhiều nghĩa:
 - bank = ngân hàng (tổ chức tài chính)
 - bank = bờ sông

Sau khi gán nhãn WordNet:

bank → bank%1:17:00:: (bờ sông)

→ Nghĩa “ngân hàng” (bank%1:14:00::) bị loại bỏ và không còn mơ hồ.

Ví dụ 2: từ plant (đa nghĩa)

The workers are building a new plant in the city.

- plant có thể là: cây cối, nhà máy
- Gán nhãn WordNet:

plant → plant%1:06:00:: (nhà máy, cơ sở sản xuất)

→ Nghĩa “thực vật” (plant%1:03:00::) bị loại bỏ.





3. Gán nhãn nghĩa của từ

3.2. Gán nhãn thủ công



word	lemma	pos	type	key	Định nghĩa theo wordnet
Ngan	person	NOUN	instance	person%1:03:00::	một con người
,	,	,	wf		
Nguyen	person	NOUN	instance	person%1:03:00::	một con người
and	and	CONJ	wf		
Thao	person	NOUN	instance	person%1:03:00::	một con người
are	be	VERB	instance	be%2:42:03::	có tính chất "là"; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
students	student	NOUN	instance	student%1:18:00::	người học đang theo học tại một cơ sở giáo dục
at	at	ADP	wf		
UIT	group	NOUN	instance	group%1:03:00::	bất kỳ số lượng thực thể (thành viên) được xem như một đơn vị
.	.	.	wf		





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
they	they	PRON	wf		
are	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
friends	friend	NOUN	wf	friend%1:18:00::	một người bạn mà bạn quen thân và tin tưởng
.	.	.	wf		
Nguyen	person	NOUN	instance	person%1:03:00::	một con người
is	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
from	from	ADP	wf		
Gia Lai	location	NOUN	instance	location%1:03:00::	một điểm hoặc vị trí trong không gian
city	city	NOUN	wf	city%1:15:00::	một khu vực đô thị lớn và đông dân; có thể gồm nhiều quận hành chính độc lập
,	,	,	wf		





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
Ngan	person	NOUN	instance	person%1:03:00::	một con người
and	and	CONJ	wf		
Thao	person	NOUN	instance	person%1:03:00::	một con người
were	be	VERB	wf		
born	bear	VERB	instance	bear%2:29:01::	khiến cho được sinh ra
in	in	ADP	wf		
coastal	coastal	ADJ	instance	coastal%3:00:00::	nằm trên, gần hoặc giáp với bờ biển
city	city	NOUN	wf	city%1:15:00::	một khu vực đô thị lớn và đông dân; có thể gồm nhiều quận hành chính độc lập
of	of	ADJ	wf		
Vung Tau	location	NOUN	instance	location%1:03:00::	một điểm hoặc vị trí trong không gian
.	.	.	wf		





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
Gia Lai	location	NOUN	instance	location%1:03:00::	một điểm hoặc vị trí trong không gian
is	be	VERB	instance	be%2:42:03::	có tính chất "là"; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
lush	lush	ADJ	instance	lush%5:00:00:abundant:00	được sản xuất hoặc mọc lên một cách vô cùng phong phú
.	.	.	wf		
the	the	DET	wf		
beach	beach	NOUN	instance	beach%1:17:00::	một vùng cát thoải xuống mép nước của biển hoặc hồ
in	in	ADP	wf		
Vung Tau	location	NOUN	instance	location%1:03:00::	một điểm hoặc vị trí trong không gian
is	be	VERB	instance	be%2:42:03::	có tính chất "là"; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
beautiful	beautiful	ADJ	instance	beautiful%3:00:00::	làm vui giác quan hoặc khơi gợi sự ngưỡng mộ trí tuệ/cảm xúc





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
and	and	CONJ	wf		
refreshing	refreshing	ADJ	instance	refreshing%5:00:00	truyền năng lượng và sức sống
.	.	.	wf		
everyone	everyone	NOUN	wf		
loves	love	VERB	instance	love%2:37:00::	có tình cảm hoặc sự thích thú mạnh mẽ đối với
this	this	DET	wf		
beach	beach	NOUN	instance	beach%1:17:00::	một vùng cát thoải xuống mép nước của biển hoặc hồ
.	.	.	wf		
Thao	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
dad	dad	NOUN	wf	dad%1:18:00::	cách nói không trang trọng chỉ “cha”, có lẽ bắt nguồn từ tiếng nói trẻ nhỏ
is	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
a	a	DET	wf		
fisherman	fisherman	NOUN	instance	fisherman%1:18:00::	người có nghề đánh bắt cá
,	,	,	wf		
and	and	CONJ	wf		
Thao	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		
mom	mom	NOUN	wf	mom%1:18:00::	cách nói không trang trọng chỉ “mẹ”
is	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
is	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
housewife	housewife	NOUN	instance	housewife%1:18:00::	người vợ quản lý gia đình trong khi chồng kiếm thu nhập
.	.	.	wf		
Ngan	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		
dad	dad	NOUN	instance	dad%1:18:00::	cách nói không trang trọng chỉ “cha”
and	and	CONJ	wf		
Ngan	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		
mom	mom	NOUN	instance	mom%1:18:00::	cách nói không trang trọng chỉ “mẹ”





3. Gán nhãn nghĩa của từ



3.2. Gán nhãn thủ công

word	lemma	pos	type	key	Định nghĩa theo wordnet
are	be	VERB	instance	be%2:42:03::	có tính chất “là”; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
fishmongers	fishmonger	NOUN	instance	fishmonger%1:18:00::	người bán cá
.	.	.	wf		
Nguyen	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		
dad	dad	NOUN	instance	dad%1:18:00::	cách nói không trang trọng chỉ “cha”
and	and	CONJ	wf		
Nguyen	person	NOUN	instance	person%1:03:00::	một con người
's	's	PRT	wf		
mom	mom	NOUN	instance	mom%1:18:00::	cách nói không trang trọng chỉ “mẹ”





3. Gán nhãn nghĩa của từ

3.2. Gán nhãn thủ công



word	lemma	pos	type	key	Định nghĩa theo wordnet
are	be	VERB	instance	be%2:42:03::	có tính chất "là"; (động từ nối, dùng với tính từ hoặc danh từ vị ngữ)
farmers	farmer	NOUN	instance	farmer%1:18:00::	người làm nghề nông / nông dân
.	.	.	wf		
they	they	PRON	wf		
love	love	VERB	instance	love%2:37:00::	có tình cảm hoặc thích rất nhiều
their	they	PRON	wf		
hometown	hometown	NOUN	instance	hometown%1:15:00::	thị trấn (hoặc thành phố) nơi bạn lớn lên hoặc nơi bạn sống chủ yếu
.	.	.	wf		





3. Gán nhãn nghĩa của từ

3.3. Tiền xử lý dữ liệu

a) Giới thiệu bộ dữ liệu

- Dữ liệu huấn luyện (Training)

- **SemCor (Miller et al., 1994)**: Corpus chú giải thủ công theo WordNet gồm 352 tài liệu và 226,040 nhãn.
- **OMSTI (Taghipour and Ng, 2015a)**: gồm hơn 1 triệu câu.

- Dữ liệu đánh giá (Test)

- Senseval-2, Senseval-3 task 1, SemEval-07 task 17, SemEval-13 task 12, SemEval-15 task 13

Thẻ	Ý nghĩa	Ví dụ
<corpus>	Tập dữ liệu gốc	<corpus lang="en" source="semcor">
<text>	Một văn bản	<text id="d000" source="br-e30">
<sentence>	Một câu trong văn bản	<sentence id="d000.s000">
<wf>	Từ không cần gán nhãn nghĩa	<wf lemma="how" pos="ADV">How</wf>
<instance>	Từ cần gán nhãn nghĩa	<instance id="d000.s000.t000" lemma="long" pos="ADJ">long</instance>

Dữ liệu trong SemCor và OMSTI được lưu dưới dạng XML





3. Gán nhãn nghĩa của từ

3.3. Tiền xử lý dữ liệu

b) Xử lý dữ liệu



Đọc và phân tích
XML



Lọc từ khóa theo
đoạn văn được chọn



Gán nhãn nghĩa từ
(Sense Key)



Kết luận

Phân tách từng văn bản
<text>
Lấy thông tin sentence
và token
=> Tổng cộng: 802.443
token

Danh sách từ ban đầu gồm 25 từ.
Loại bỏ stopwords => 18 từ chính
Giữ các câu có lemma có trong
danh sách.
=> Số câu: 1.024.
Lọc các câu quá dài (số token >
50) => còn lại 922 câu.

Dữ liệu SemCor/OMSTI đi
kèm file gold.key.txt, định
nghĩa nhãn nghĩa từ cho từng
instance: Mỗi dòng: token_id
sense_key
Sử dụng token_id để map
nhãn sense vào DataFrame.

- Không thiếu từ khóa quan trọng.
- Đa dạng về POS và từ loại (nouns, verbs, adjectives).
- Độ dài câu hợp lý, phù hợp cho các mô hình học sâu.

source	text_id	sentence_id	token_id	word	lemma	pos	type
semcor	d000	d000.s000	None	How	how	ADV	wf
semcor	d000	d000.s000	d000.s000.t000	long	long	ADJ	instance



word	lemma	pos	type	key
Greene	person	NOUN	instance	person%1:03:00::
was	be	VERB	instance	be%2:42:03::





3. Gán nhãn nghĩa của từ

3.3. Tiền xử lý dữ liệu

c) Phân tích dữ liệu

Dữ liệu huấn luyện

Số câu: 922

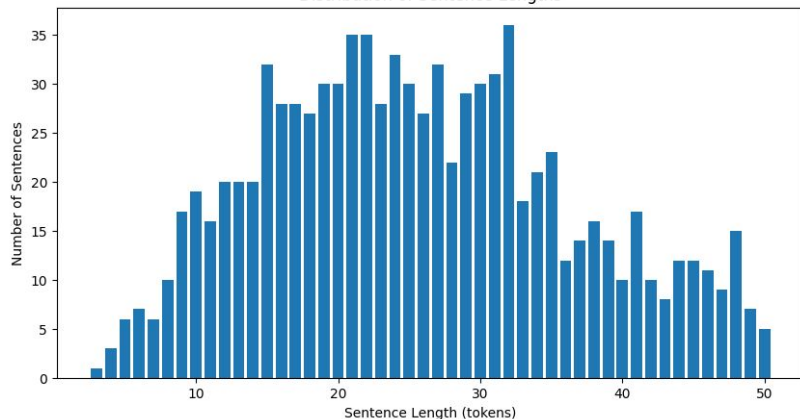
Số token: 23.683

Số token có nhãn nghĩa (instance): 5.726

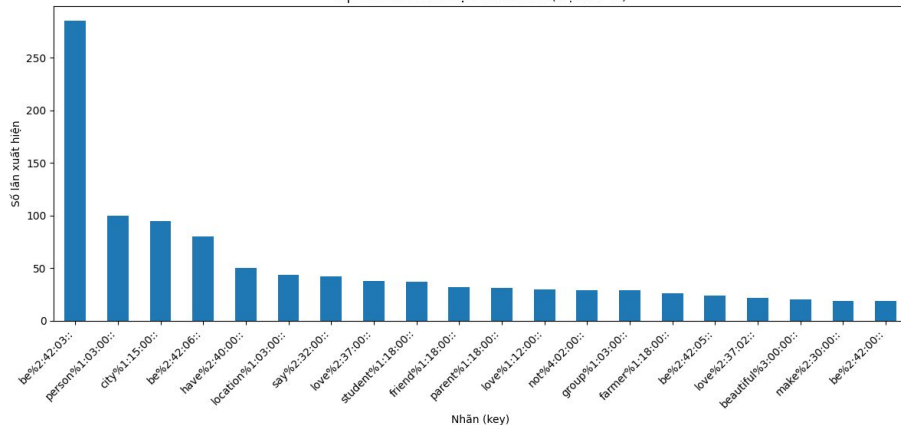
Số token bình thường (wf): 17.957

Tổng số nhãn (key) cần dự đoán: 3.047

Distribution of Sentence Lengths



Top 20 nhãn xuất hiện nhiều nhất (loại trừ 'O')



Sentence	['Greene', 'was', 'in', 'actuality', 'a', 'young', 'ruffian', 'from', 'Kent', ',', ...]
Lemma	['person', 'be', 'in', 'actuality', 'a', 'young', 'ruffian', 'from', 'kent', ',', ...]
Labels	['person%1:03:00::', 'be%2:42:03::', 'O', 'actuality%1:26:00::', 'O', 'young%3:00:00::', ...]





3. Gán nhãn nghĩa của từ

3.3. Tiền xử lý dữ liệu

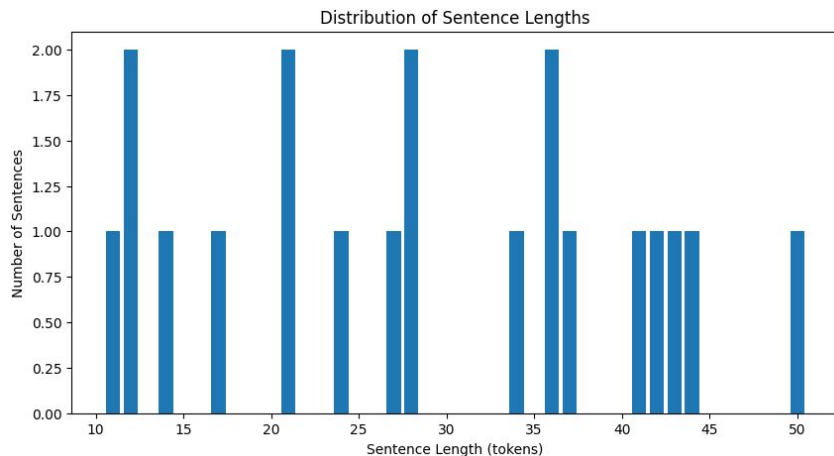
c) Phân tích dữ liệu

Dữ liệu đánh giá

Số câu: 21

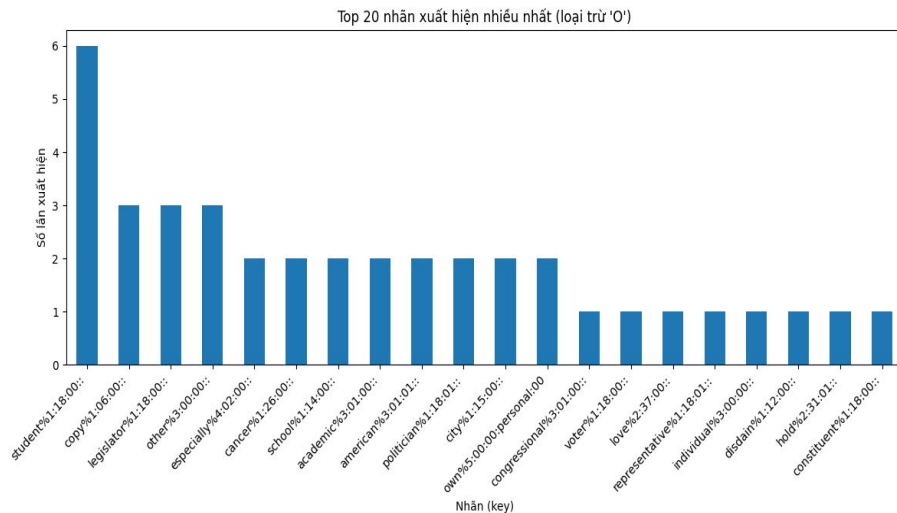
Số token khác nhau: 331

Tổng số nhãn (key) dự đoán: 152



Số token có nhãn nghĩa (instance): 173

Số token bình thường (wf): 158





3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình



Mô hình được sử dụng:

- **Most Frequent Sense (MFS):** Thường được dùng là baseline. Gán cho từ nghĩa phổ biến nhất trong WordNet dựa trên dữ liệu huấn luyện.

Ví dụ : Từ bank

Các nghĩa trong WordNet (rút gọn):

bank%1:14:00:: ngân hàng tài chính

bank%1:17:00:: bờ sông

Giả sử trong tập huấn luyện:

bank%1:14:00:: xuất hiện 75%

bank%1:17:00:: xuất hiện 25%

Most Frequent Sense: bank%1:14:00::

Áp dụng MFS:

I deposited money in the bank -> bank%1:14:00::

The fisherman sat on the bank of the river -> bank%1:14:00::

Dù câu thứ hai nói về bờ sông, MFS vẫn gán ngân hàng.

Ví dụ : Từ interest

Các nghĩa phổ biến:

interest%1:09:00:: sự quan tâm

interest%1:21:00:: lãi suất

Giả sử trong tập huấn luyện:

interest (sự quan tâm): 60%

interest (lãi suất): 40%

MFS = interest%1:09:00::

Áp dụng MFS:

She showed interest in art -> interest%1:09:00::

The bank raised interest rates-> interest%1:09:00::





3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Mô hình được sử dụng:

- **Most Frequent Sense (MFS):** Gán cho từ nghĩa phổ biến nhất trong WordNet dựa trên dữ liệu huấn luyện. Thường được dùng là baseline.

lemma2mfs

```
{'person': 'person%1:03:00::',  
  'be': 'be%2:42:03::',  
  'actuality': 'actuality%1:26:00::',  
  'young': 'young%3:00:00::',  
  'ruffian': 'ruffian%1:18:00::',  
  'kent': 'kent%1:15:00::',  
  'break_with': 'break_with%2:41:00::',  
  'parent': 'parent%1:18:00::',  
  'keep_company': 'keep_company%2:42:00::',  
  'prefer': 'prefer%2:37:00::',  
  'pimp': 'pimp%1:18:00::',  
  'pander': 'pander%1:18:00::',  
  'whore': 'whore%1:18:00::',  
  'born': 'born%3:00:00::',  
  'prosperous': 'prosperous%5:00:00:rich:00',  
  'family': 'family%1:14:02::',  
  'carefree': 'carefree%5:00:00:untroubled:00',  
  'indolence': 'indolence%1:07:00::',
```

```
'see': 'see%2:39:00::',  
'last': 'last%3:00:00::',  
'reflection': 'reflection%1:10:01::',  
'old': 'old%3:00:01::',  
'happy': 'happy%3:00:00::',  
'existence': 'existence%1:26:00::',  
'reason': 'reason%1:16:00::',  
'love': 'love%2:37:00::',  
'different': 'different%3:00:00::',  
'volume': 'volume%1:06:00::',  
'history': 'history%1:10:00::',  
'intend': 'intend%2:31:01::',  
'general': 'general%3:00:00::',  
'particularly': 'particularly%4:02:01::',  
'suited': 'suited%5:00:00:fit:02',  
'high_school': 'high_school%1:14:00::',  
'student': 'student%1:18:00::',  
'bear': 'bear%2:29:01::',  
'small': 'small%3:00:00::',  
'danish': 'danish%3:01:00::',  
'town': 'town%1:15:00::',  
'location': 'location%1:03:00::',  
'island': 'island%1:17:00::',  
'south-central': 'south-central%5:00:00:south:00',  
'part': 'part%1:15:00::',  
'denmark': 'denmark%1:15:00::',  
'august': 'august%1:28:00::',  
'fourteenth': 'fourteenth%5:00:00:ordinal:00',  
'early': 'early%3:00:02::',  
'literature': 'literature%1:09:00::',  
'remain': 'remain%2:30:00::',  
'entire': 'entire%5:00:01:whole:00',  
'life': 'life%1:09:00::',  
'also': 'also%4:02:00::',  
'meet': 'meet%2:41:01::',  
'then': 'then%4:02:00::',
```



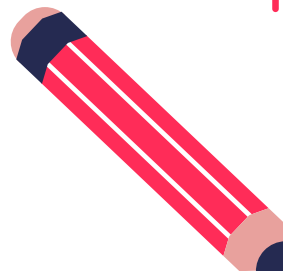


3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Mô hình được sử dụng:

- **BERT**: Dùng BERT để dự đoán key WordNet.
- **RoBERTa**: Phiên bản cải tiến của BERT, cho kết quả WSD chính xác hơn.



Trích xuất đặc trưng với BERT

Bước 1: Token hoá

Sử dụng BERT tokenizer

Bước 2: Gán nhãn token

- Từ không tách: giữ nguyên nhãn WordNet
- Từ bị tách: Subword đầu: giữ nhãn thật, Subword sau: gán nhãn X

Ví dụ: actual | ##ity → actuality%1:26:00:: | X

Bước 3: Mã hoá bằng BERT

Mỗi token → vector ngữ cảnh 768 chiều

Bước 4: Chọn vector đại diện

Chỉ dùng vector của subword đầu

Token nhãn X: Không tính loss

Mỗi từ được biểu diễn bởi một vector ngữ cảnh duy nhất

Classification Head (BertForTokenClassification)

Bước 1: Đầu ra từ BERT

Chuỗi vector ngữ cảnh:

$$H=[h_1, h_2, \dots, h_n], h_i \in R^{768}$$

Bước 2: Lớp tuyến tính (Fully-connected layer)

$R^{768} \rightarrow R^{3036}$ (Ánh xạ vector token → logits của các sense WordNet)

Bước 3: Softmax & dự đoán

Áp dụng Softmax

Chọn nhãn có xác suất cao nhất

→ Mỗi token được gán một WordNet sense





3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

- **MFS:** Chỉ số rất thấp. Recall cao hơn Precision một chút ở Macro, cho thấy nó nắm bắt được một số trường hợp phổ biến, nhưng nhìn chung là không tốt.
- **BERT-Base, BERT-Large:** Đã gần như hoàn hảo nhưng vẫn có lỗi nhỏ.
- **RoBERTa:** thể hiện mức quá khớp cao nhất, đạt Macro Precision tuyệt đối (1.0000), nghĩa là nó không có một lỗi sai dương tính giả (False Positive) nào trên bất kỳ nhãn nào trong tập train.

Mô hình	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Micro f1	Macro f1	Accuracy
Most Frequency Sense	0.7669	0.5981	0.7669	0.7800	0.7669	0.6457	0.7669
BERT (bert-base-uncased)	0.9959	0.9683	0.9959	0.9722	0.9959	0.9695	0.9959
BERT (bert-large-uncased)	0.9994	0.9973	0.9994	0.9969	0.9994	0.9971	0.9994
RoBERTa	0.9999	1	0.9999	0.9999	0.9999	0.9999	0.9999





3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Đánh giá mô hình trên tập test **WSD_Unified_Evaluation_Datasets**:

- **MFS**: MFS có xu hướng dự đoán chính xác hơn là dự đoán được các nhãn hiếm.
- **BERT-Base**: Đạt Micro F1 rất cao và có Macro Recall (0.2226) tốt hơn Macro Precision, cho thấy xu hướng bao phủ các nhãn hiếm tốt hơn một chút.
- **BERT-Large**: Macro F1 thấp nhất trong nhóm BERT => tập trung quá mức vào các nhãn phổ biến
- **RoBERTa**: Cân bằng Macro khá tốt, nhưng Micro F1 thấp hơn BERT.

Mô hình	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Micro f1	Macro f1	Accuracy
Most Frequency Sense	0.6263	0.1753	0.6263	0.1939	0.6263	0.1811	0.6263
BERT (bert-base-uncased)	0.7699	0.2129	0.7699	0.2226	0.7699	0.211	0.7699
BERT (bert-large-uncased)	0.7837	0.1874	0.7837	0.191	0.7837	0.1845	0.7837
RoBERTa	0.7907	0.1673	0.7907	0.1759	0.7907	0.1656	0.7907





3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Đánh giá mô hình trên những câu trong đoạn văn bản nhóm tạo:

- **MFS:** Đạt cao nhất tại mọi chỉ số,, cho thấy việc đoán nghĩa phổ biến nhất lại hiệu quả hơn.
- **BERT-Base:** Micro F1 thấp nhất, Macro Recall cực thấp. Quá trình overfit đã dẫn đến thất bại nghiêm trọng trên ngữ cảnh mới.
- **BERT-Large:** Mặc dù Micro F1 thấp hơn MFS, nhưng Macro F1 cao nhất (0.3374), cho thấy nó xử lý các nhãn hiếm trong ngữ cảnh tùy chỉnh tốt hơn các phiên bản khác.
- **RoBERTa:** Micro F1 cao thứ hai (chỉ sau MFS), và Macro F1 rất cao, gần bằng MFS. RoBERTa cho thấy khả năng tổng quát hóa tốt nhất trên tập test so với các mô hình BERT khác.

Mô hình	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Micro f1	Macro f1	Accuracy
Most Frequency Sense	0.6292	0.5824	0.6292	0.5979	0.6292	0.5882	0.6292
BERT (bert-base-uncased)	0.4435	0.2688	0.4435	0.1858	0.4435	0.2062	0.4435
BERT (bert-large-uncased)	0.4522	0.4029	0.4522	0.3118	0.4522	0.3374	0.4522
RoBERTa	0.5041	0.4387	0.5041	0.3508	0.5041	0.3816	0.5041



3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Câu: Ngan , Nguyen and Thao are students at UIT .

Ngan	lemma: ngan	GOLD: person%1:03:00::	PRED: 0	X
,	lemma: ,	GOLD: 0	PRED: 0	✓
Nguyen	lemma: nguyen	GOLD: person%1:03:00::	PRED: 0	X
and	lemma: and	GOLD: 0	PRED: 0	✓
Thao	lemma: thao	GOLD: person%1:03:00::	PRED: 0	X
are	lemma: are	GOLD: be%2:42:03::	PRED: 0	X
students	lemma: student	GOLD: student%1:18:00::	PRED: student%1:18:00::	✓
at	lemma: at	GOLD: 0	PRED: 0	✓
UIT	lemma: uit	GOLD: group%1:03:00::	PRED: 0	X
.	lemma: .	GOLD: 0	PRED: 0	✓

Câu: they are friends .

they	lemma: they	GOLD: 0	PRED: 0	✓
are	lemma: are	GOLD: be%2:42:03::	PRED: 0	X
friends	lemma: friend	GOLD: friend%1:18:00::	PRED: friend%1:18:00::	✓
.	lemma: .	GOLD: 0	PRED: 0	✓

Most Frequency Sense

- Danh từ riêng (Ngan, Nguyen, Thao, UIT) không có lemma chuẩn trong WordNet nên bị gán nhãn sai.
- Không nhận biết được are là động từ be
- Một số từ thường (students, friends, they, dấu câu) được gán nhãn đúng.

3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Câu: Nguyen is from Gialai city ,

Nguyen	lemma: nguyen	GOLD: person%1:03:00::	PRED: 0	X
is	lemma: is	GOLD: be%2:42:03::	PRED: 0	X
from	lemma: from	GOLD: 0	PRED: 0	✓
Gialai	lemma: gialai	GOLD: location%1:03:00::	PRED: 0	X
city	lemma: city	GOLD: city%1:15:00::	PRED: city%1:15:00::	✓
,	lemma: ,	GOLD: 0	PRED: 0	✓

Câu: Ngan and Thao were born in coastal city of Vungtau .

Ngan	lemma: ngan	GOLD: person%1:03:00::	PRED: 0	X
and	lemma: and	GOLD: 0	PRED: 0	✓
Thao	lemma: thao	GOLD: person%1:03:00::	PRED: 0	X
were	lemma: were	GOLD: 0	PRED: 0	✓
born	lemma: born	GOLD: bear%2:29:01::	PRED: born%3:00:00::	X
in	lemma: in	GOLD: 0	PRED: in%4:02:01::	X
coastal	lemma: coastal	GOLD: coastal%3:00:00::	PRED: coastal%3:01:00::	X
city	lemma: city	GOLD: city%1:15:00::	PRED: city%1:15:00::	✓
of	lemma: of	GOLD: 0	PRED: 0	✓
Vungtau	lemma: vungtau	GOLD: location%1:03:00::	PRED: 0	X
.	lemma: .	GOLD: 0	PRED: 0	✓

Most Frequency Sense

- PP MFS dự đoán tốt các từ phổ biến như danh từ chung city, giới từ và dấu câu, nhưng thường sai với danh từ riêng Việt Nam do WordNet không có sense phù hợp.
- Các động từ (were, born) và tính từ (coastal) cũng dễ bị gán sai vì PP MFS luôn chọn sense phổ biến nhất, vốn không phù hợp với ngữ cảnh cụ thể, trong khi các từ phổ biến vẫn được dự đoán chính xác.

3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Sentence: Ngan , Nguyen and Thao are students at UIT .

Word	True Label	Pred Label
Ngan	person%1:03:00::	0
,	0	0
Nguyen	person%1:03:00::	0
and	0	0
Thao	person%1:03:00::	0
are	be%2:42:03::	be%2:42:07::
students	student%1:18:00::	0
at	0	0
UIT	group%1:03:00::	0
.	0	0

Sentence: they are friends .

Word	True Label	Pred Label
they	0	0
are	be%2:42:03::	be%2:42:03::
friends	friend%1:18:00::	friend%1:18:00::
.	0	0

RoBERTa

- Câu 1: RoBERTa xử lý tốt từ chức năng và dấu câu, nhưng gặp khó khăn với danh từ riêng tiếng Việt và một số danh từ chung, dù nhận diện đúng động từ are.
- Câu 2: RoBERTa dự đoán chính xác tất cả các từ phổ biến, cho thấy hiệu quả cao với các sense WordNet rõ ràng.



3. Gán nhãn nghĩa của từ

3.4. Huấn luyện mô hình

Sentence: Nguyen is from Gialai city ,

Word	True Label	Pred Label
------	------------	------------

Nguyen	person%1:03:00::	0
is	be%2:42:03::	be%2:42:05::
from	0	0
Gialai	location%1:03:00::	0
city	city%1:15:00::	0
,	0	0

Sentence: Ngan and Thao were born in coastal city of Vungtau .

Word	True Label	Pred Label
------	------------	------------

Ngan	person%1:03:00::	person%1:03:00::
and	0	0
Thao	person%1:03:00::	person%1:03:00::
were	0	0
born	bear%2:29:01::	bear%2:29:01::
in	0	0
coastal	coastal%3:00:00::	coastal%3:00:00::
city	city%1:15:00::	city%1:15:00::
of	0	0
Vungtau	location%1:03:00::	location%1:03:00::
.	0	0



RoBERTa

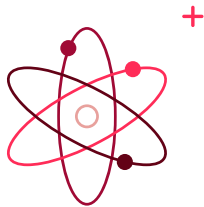
- Câu 3: RoBERTa dự đoán đúng các từ phổ biến và dấu câu nhưng chưa nhận diện tốt danh từ riêng tiếng Việt và còn gán sai sense WordNet cho một số từ.
- Câu 4: Ở trường hợp còn lại, mô hình nhận diện chính xác cả danh từ riêng địa phương lẫn từ phổ biến và gán sense WordNet phù hợp.



04

THỰC NGHIỆM

- Cài đặt Prolog
- Xử lý câu truy vấn
- Ứng dụng





4. THỰC NGHIỆM



4.1. Cài đặt Prolog

1. Xây dựng Cơ sở Tri thức Lỗi

Câu 1: Ngan, Nguyen, and Thao are students at UIT.

- Logic: $F_1 = (\text{student}(\text{Ngan}) \wedge \text{at}(\text{Ngan}, \text{UIT})) \wedge (\text{student}(\text{Thao}) \wedge \text{at}(\text{Thao}, \text{UIT})) \wedge (\text{student}(\text{Nguyen}) \wedge \text{at}(\text{Nguyen}, \text{UIT}))$
- Prolog: `student(ngan). student(thao). student(nguyen).at(ngan, uit). at(thao, uit). at(nguyen, uit).`

Câu 2: They are friends.

- Logic: $F_2 = (\text{friend}(\text{Ngan}, \text{Thao}) \wedge \text{friend}(\text{Thao}, \text{Ngan})) \wedge (\text{friend}(\text{Ngan}, \text{Nguyen}) \wedge \text{friend}(\text{Nguyen}, \text{Ngan})) \wedge (\text{friend}(\text{Thao}, \text{Nguyen}) \wedge \text{friend}(\text{Nguyen}, \text{Thao}))$
- Prolog: `friend(ngan, thao). friend(thao, ngan). friend(ngan, nguyen). friend(nguyen, ngan). friend(thao, nguyen). friend(nguyen, thao).`

Câu 5: The beach in Vung Tau is beautiful and refreshing.

- Logic: $F_5 = \text{beach}(\text{BH}) \wedge \text{beautiful}(\text{BH}) \wedge \text{refreshing}(\text{BH})$
- Prolog: `beautiful(X) :- beach(X), in(X, vung_tau). refreshing(X) :- beach(X), in(X, vung_tau).`

Câu 6: Everyone loves Vung Tau's beaches. (Giả định 'Everyone' là 'person(y)')

- Logic: $F_6 = \forall y, \text{person}(y) \wedge \text{beach}(\text{BH}) \wedge \text{in}(\text{BH}, \text{Vung_Tau}) \rightarrow \text{loves}(y, \text{BH})$
- Prolog: `love(Y, X) :- person(Y), beach(X), in(X, vung_tau)`



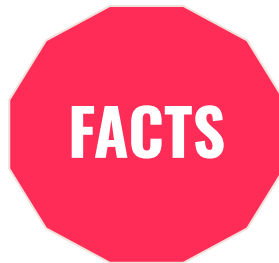


4. THỰC NGHIỆM

4.1. Cài đặt Prolog

1. Xây dựng Cơ sở Tri thức Lỗi (Core KB)

% Sinh viên ở UIT
student(ngan).
student(thao).
student(nguyen).
at(ngan, uit).
at(thao, uit).
at(nguyen, uit).
% Bạn bè
friend(ngan, thao).
friend(thao, ngan).
friend(ngan, nguyen).
friend(nguyen, ngan).
friend(thao, nguyen).
friend(nguyen, thao).



% Quê quán & thành phố
from(nguyen, gia_lai).
city(gia_lai).
city(vung_tau).
lush(gia_lai).
coastal(vung_tau).
% Nơi sinh
be_born(ngan, vung_tau).
be_born(thao, vung_tau).
% Gia đình Ngan
dad(ngan_dad).
mom(ngan_mom).
s(ngan, ngan_dad).
s(ngan, ngan_mom).
fishmonger(ngan_dad).
fishmonger(ngan_mom).



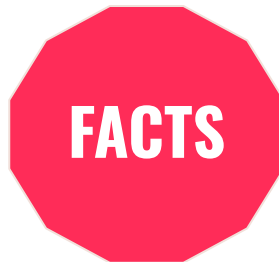


4. THỰC NGHIỆM

4.1. Cài đặt Prolog

1. Xây dựng Cơ sở Tri thức Lỗi (Core KB)

% Gia đình Nguyen
dad(nguyen_dad).
mom(nguyen_mom).
s(nguyen, nguyen_dad).
s(nguyen, nguyen_mom).
farmer(nguyen_dad).
farmer(nguyen_mom).
% Con người
person(thao).
person(ngan).
person(nguyen).
person(thao_dad).
person(thao_mom).
person(ngan_dad).



person(ngan_mom).
person(nguyen_dad).
person(nguyen_mom).
% Quan hệ gia đình
family(thao, thao_dad).
family(thao, thao_mom).
family(ngan, ngan_dad).
family(ngan, ngan_mom).
family(nguyen, nguyen_dad).
family(nguyen, nguyen_mom).
% Quê hương
hometown(nguyen, gia_lai).
hometown(ngan, vung_tau).
hometown(thao, vung_tau).



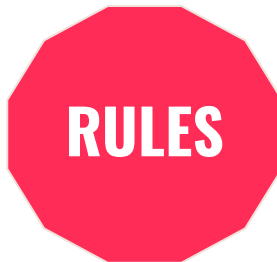


4. THỰC NGHIỆM

4.1. Cài đặt Prolog

1. Xây dựng Cơ sở Tri thức Lỗi (Core KB)

% Vai trò ở UIT
role(X, student) :-
student(X).
% Bãi biển ở VT thì đẹp và
mát mẻ
beautiful(X) :- beach(X),
in(X, vung_tau).
refreshing(X) :- beach(X),
in(X, vung_tau).
% Người thì yêu bãi biển ở
VT
love(Y, X) :- person(Y),
beach(X), in(X, vung_tau).



% Người yêu quê hương của họ
love(Y, H) :- hometown(Y, H).
% Một người thuộc gia đình F
nếu family(X,F)
in_family(X, F) :- family(X, F).
% Bãi biển ở VT
beach(beach_vt).
in(beach_vt, vung_tau).
% Gia đình Thao
dad(thao_dad).
mom(thao_mom).
s(thao, thao_dad).
s(thao, thao_mom).
fisherman(thao_dad).
housewife(thao_mom).

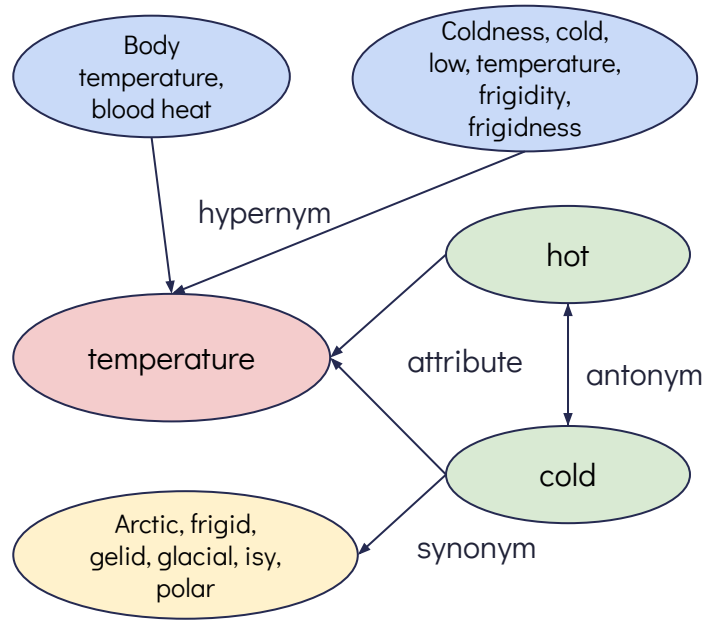




4. THỰC NGHIỆM

4.1. Cài đặt Prolog

2. Mở rộng Ngữ nghĩa bằng WordNet (WordNet Semantic Expansion)



Loại tri thức	Ý nghĩa	Vai trò
Synonyms (từ đồng nghĩa)	Các từ khác chia sẻ cùng một synset	Dùng mở rộng truy vấn hoặc phát hiện paraphrase
Antonyms (từ trái nghĩa)	Các từ có nghĩa đối lập	Hỗ trợ phân loại cảm xúc, mô hình phủ định logic
Hypernyms (siêu nghĩa)	Danh mục khái niệm bao trùm	Dùng biểu diễn phân cấp (ví dụ: fisherman → worker → person)
Hyponyms (dưới nghĩa)	Các khái niệm con cụ thể hơn	Hỗ trợ gợi ý từ vựng, suy luận đặc điểm
Meronyms / Holonyms	Quan hệ <i>bộ phận – toàn bộ</i>	Dùng liên kết thực thể trong đồ thị tri thức





4. THỰC NGHIỆM

4.1. Cài đặt Prolog

2. Mở rộng Ngữ nghĩa bằng WordNet (WordNet Semantic Expansion)

```
get_synset_from_wordnet_id("love%2:37:00::", lemma="love", pos='2')
```

```
{'synset': 'love.v.01',  
'definition': 'have a great affection or liking for',  
'synonyms': ['love'],  
'antonyms': ['hate'],  
'hypernyms': [],  
'hyponyms': ['love.v.03', 'adore.v.01', 'care_for.v.02', 'dote.v.02'],  
'member_meronyms': [],  
'member_holonyms': [],  
'part_meronyms': [],  
'part_holonyms': [],  
'substance_meronyms': [],  
'substance_holonyms': []}
```

Loại tri thức	Ý nghĩa	Vai trò
Synonyms (từ đồng nghĩa)	Các từ khác chia sẻ cùng một synset	Dùng mở rộng truy vấn hoặc phát hiện paraphrase
Antonyms (từ trái nghĩa)	Các từ có nghĩa đối lập	Hỗ trợ phân loại cảm xúc, mô hình phủ định logic
Hypernyms (siêu nghĩa)	Danh mục khái niệm bao trùm	Dùng biểu diễn phân cấp (ví dụ: fisherman → worker → person)
Hyponyms (dưới nghĩa)	Các khái niệm con cụ thể hơn	Hỗ trợ gợi ý từ vựng, suy luận đặc điểm
Meronyms / Holonyms	Quan hệ <i>bộ phận – toàn bộ</i>	Dùng liên kết thực thể trong đồ thị tri thức



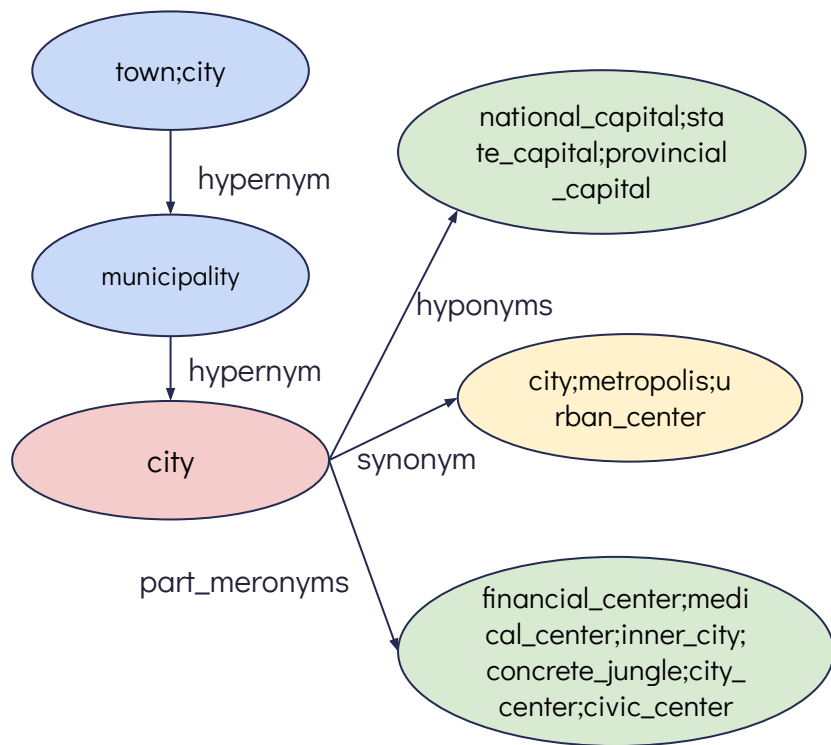


4. THỰC NGHIỆM

4.1. Cài đặt Prolog



2. Mở rộng Ngữ nghĩa bằng WordNet (WordNet Semantic Expansion)



Ngoài các **synonyms**, **antonyms**, **meronyms**, **holonyms**, khi mở rộng nghĩa từ WordNet, chúng ta còn có thể trích xuất thông tin phân cấp khái niệm thông qua các trường **hypernyms_level1**, **hypernyms_level2**, **hyponyms_level1**, **hyponyms_level2**





4. THỰC NGHIỆM

4.1. Cài đặt Prolog

2. Mở rộng Ngữ nghĩa bằng WordNet (WordNet Semantic Expansion)

Từ **key** đã được gán nhãn, trích xuất toàn bộ

quan hệ ngữ nghĩa từ Wordnet gồm:

- synonyms
- antonyms
- Hypernyms_level1
- Hypernyms_level2
- Hyponyms_level1
- hyponyms_level2
- meronyms (member, part, substance)
- holonyms (member, part, substance)

word	key	synset	definition	synonyms	antonyms	hypernyms_level1	hypernyms_level2	hyponyms_level1
Ngân	person%1:03:00::	person.n.01	a human being	person;individual;someone;somebody; mortal;soul		causal_agent.n.01;organism.n.01	engine.n.02;deu	gainer.n.01;birth.n.05;misogamis
Nguyen	person%1:03:00::	person.n.01	a human being	person;individual;someone;somebody; mortal;soul		causal_agent.n.01;organism.n.01	engine.n.02;deu	gainer.n.01;birth.n.05;misogamis
Thao	person%1:03:00::	person.n.01	a human being	person;individual;someone;somebody; mortal;soul		causal_agent.n.01;organism.n.01	engine.n.02;deu	gainer.n.01;birth.n.05;misogamis
are	be%2:42:03::	be.v.01	have the quality of being; (copula, used with an adjective or a predicate noun)	be				come_in_handy.v.01;want.v.02;li
students	student%1:18:00::	student.n.01	a learner who is enrolled in an educational institution	student;pupil;educatee		enrollee.n.01	student.n.01	law_student.n.01;withdrawer.n.01
UIT	group%1:03:00::	group.n.01	any number of entities (members) considered as a unit	group;grouping		abstraction.n.06	communication.n.08;association.n.08;circuit.n.05;rare	come_in_handy.v.01;want.v.02;li
are	be%2:42:03::	be.v.01	have the quality of being; (copula, used with an adjective or a predicate noun)	be				come_in_handy.v.01;want.v.02;li
friends	friend%1:18:00::	friend.n.01	a person you know well and regard with affection and trust	friend		person.n.01	gainer.n.01;birth	mate.n.08;girlfriend.n.01;confidant
Nguyen	person%1:03:00::	person.n.01	a human being	person;individual;someone;somebody; mortal;soul		causal_agent.n.01;organism.n.01	engine.n.02;deu	gainer.n.01;birth.n.05;misogamis
is	be%2:42:03::	be.v.01	have the quality of being; (copula, used with an adjective or a predicate noun)	be				come_in_handy.v.01;want.v.02;li
Cia-let	location%1:03:00::	location.n.01	a point or extent in space	location		abstraction.n.01	ship.n.03;day.n.03;post.n.05;southwest	



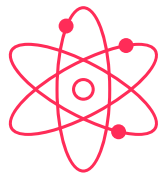
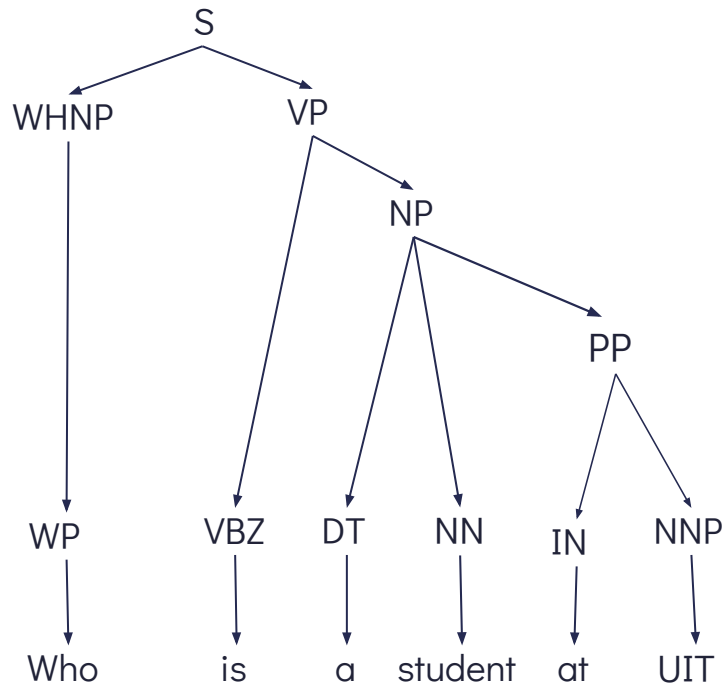


4. THỰC NGHIỆM

4.2. Xử lý câu truy vấn

Biến đổi câu truy vấn sang FOL bằng biểu thức lambda

“Who is a student at UIT?”

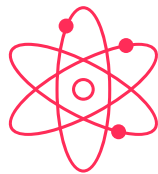




4. THỰC NGHIỆM

4.2. Xử lý câu truy vấn

Biến đổi câu truy vấn sang FOL bằng biểu thức lambda



Biểu thức lambda

NNP ($\lambda P.(P @ UIT)$) \rightarrow UIT

IN ($\lambda Q.\lambda X.(Q @ \lambda Y.at(X,Y))$) \rightarrow at

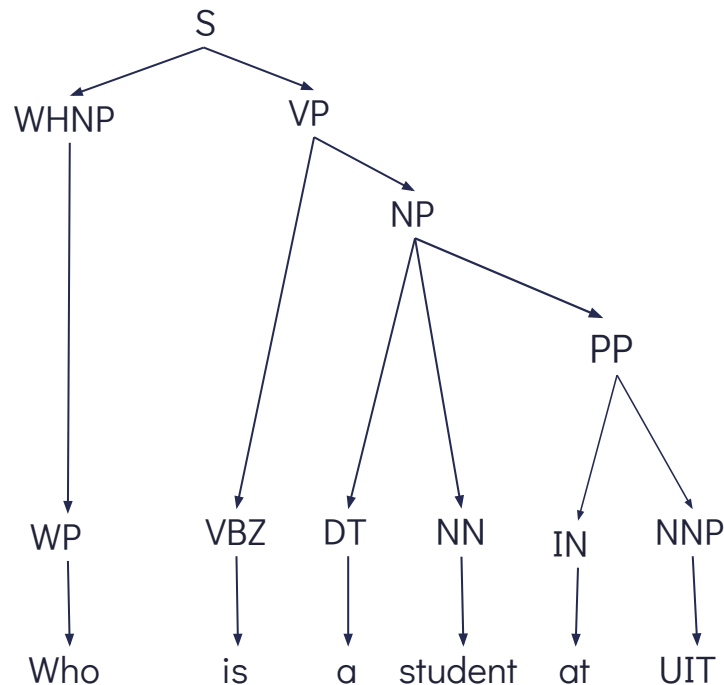
NN ($\lambda P.student(P)$) \rightarrow student

DT ($\lambda P.\lambda Q.\exists X.(P @ X \wedge Q @ X)$) \rightarrow

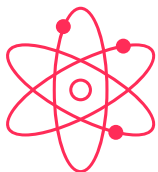
a

VBZ ($\lambda P.P$) \rightarrow is

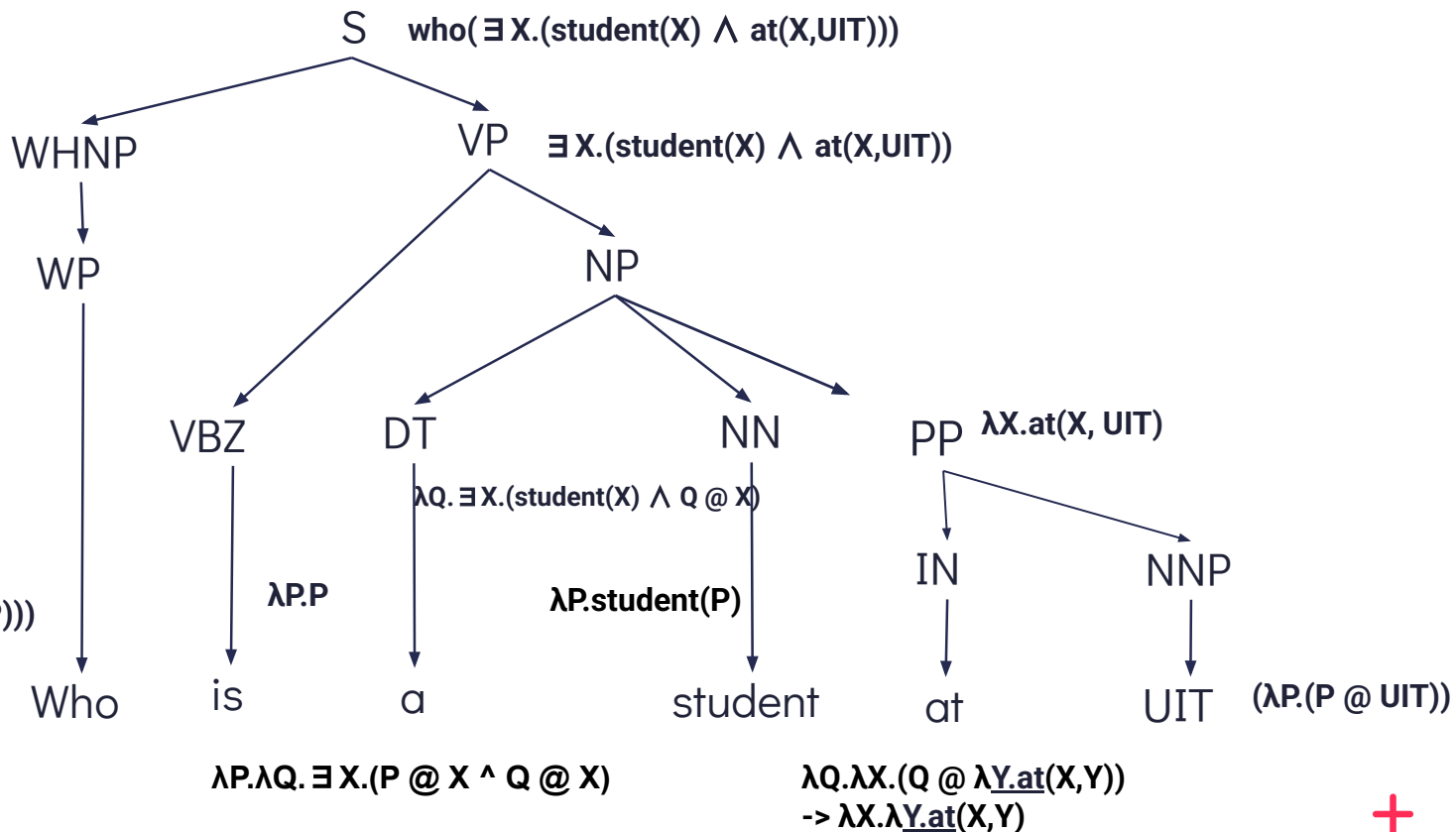
WP ($\lambda P.(who(P))$) \rightarrow who



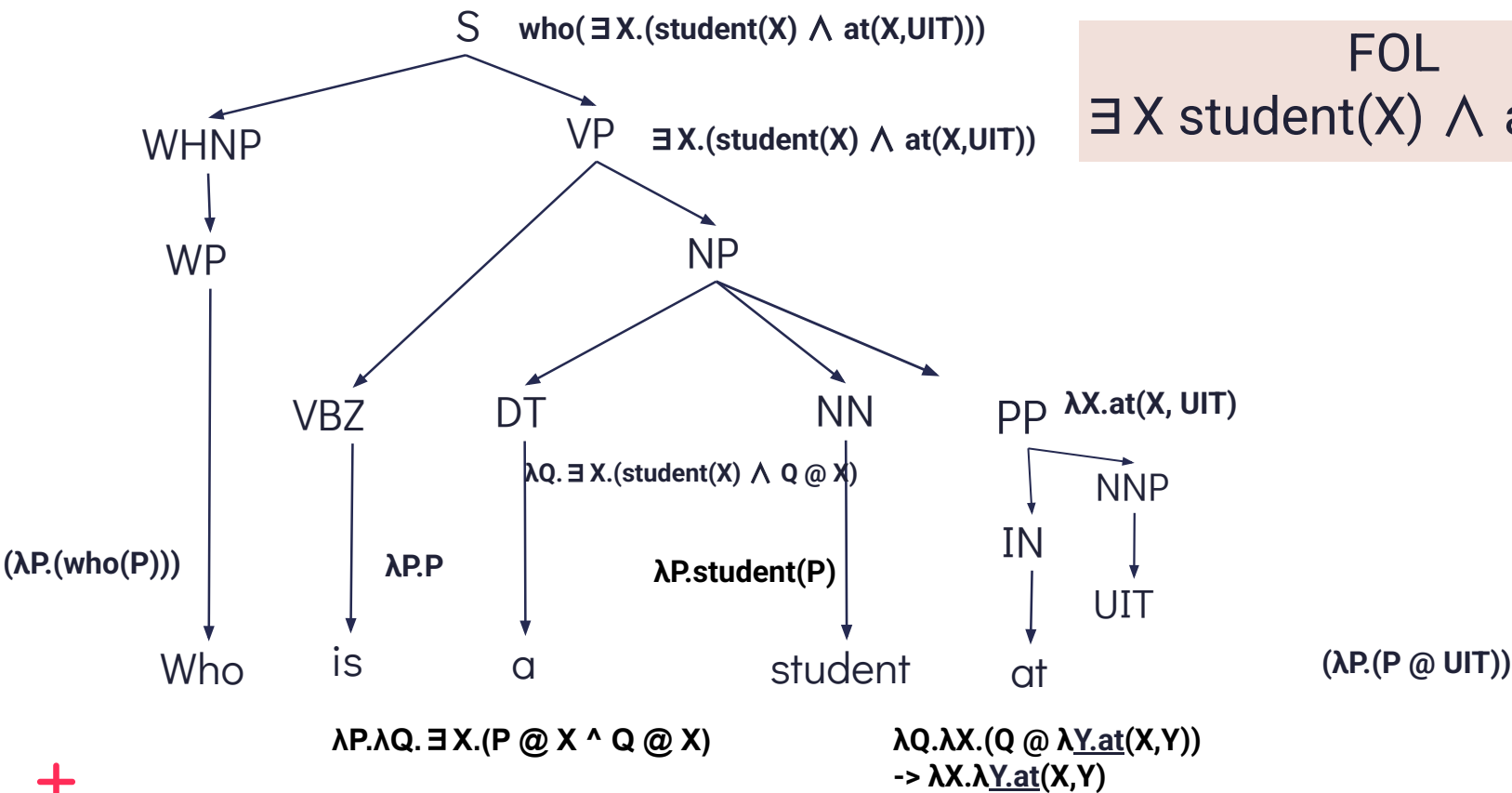
Tính toán
biểu thức
lambda



$(\lambda P.(\text{who}(P)))$



Biến đổi câu truy vấn sang FOL bằng biểu thức lambda

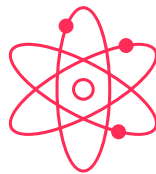


FOL
 $\exists X \text{ student}(X) \wedge \text{at}(X, \text{UIT})$



4. THỰC NGHIỆM

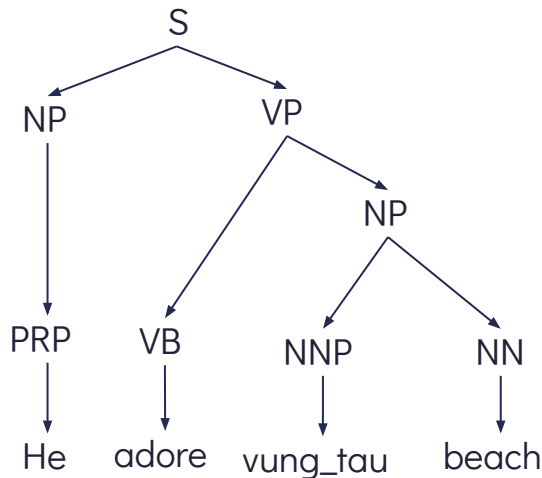
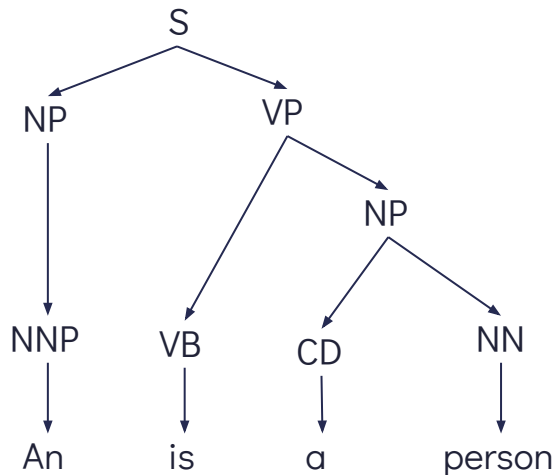
4.2. Xử lý câu truy vấn



Biến đổi câu truy vấn sang FOL bằng DRS

“An is a person. He adore Vung Tau beach?”

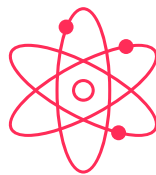
1. Tập $D = \{d\}$
2. Tập $C = \{\text{“An is a person”, “He adore Vung Tau beach”}\}$
3. Tập câu cú pháp





4. THỰC NGHIỆM

4.2. Xử lý câu truy vấn

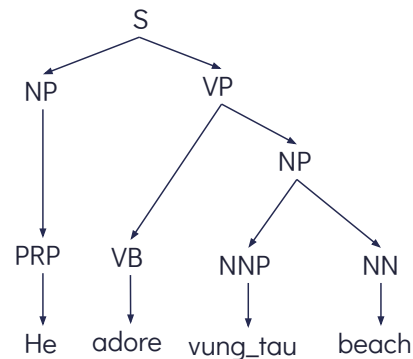
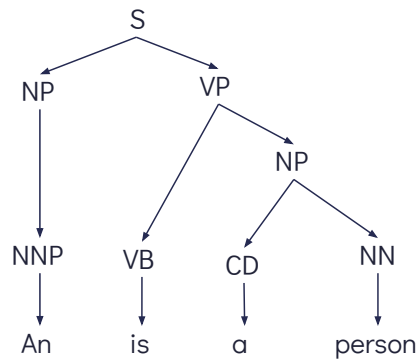


Biến đổi câu truy vấn sang FOL bằng DRS

"An is a person. He adore Vung Tau beach"

4. Tập T chứa DRS từ

1. NNP ($\lambda P.\langle\{an\},\{\}\rangle^{\otimes} (P@an)) \rightarrow An$
2. VB ($_$) $\rightarrow is$
3. CD ($\lambda P.\lambda Q.\langle\{X\},\{\}\rangle^{\otimes} (P@X)^{\otimes} (Q@X)) \rightarrow a$
4. NN ($\lambda X.\langle\{\},\{person(X)\}\rangle$) $\rightarrow person$
5. PRP ($\lambda P.a_x\langle\{X\},\{\}\rangle^{\otimes} (P@X)) \rightarrow he$
6. VB ($\lambda P.\lambda X.P@ Y.\langle\{\},\{adore(X, Y)\}\rangle$) $\rightarrow adore$
7. NNP ($\lambda P.\langle\{vung_tau\},\{\}\rangle^{\otimes} (P@vung_tau)) \rightarrow Vung\ Tau$
8. NN ($\lambda X.\langle\{\},\{beach(X)\}\rangle$) $\rightarrow beach$



5. Tập G chứa các quy tắc hợp thành DRS

$S(DN @DV) \rightarrow NP(DN) VP(DV)$
 $NP(DN) \rightarrow NNP(DN)$
 $NP(DN) \rightarrow PRP(DN)$

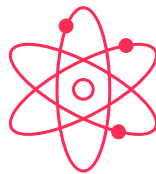
$VP(DV@DN) \rightarrow VB(DV) NP(DN)$
 $NP(DC@DN) \rightarrow CD(DC) NN(DN)$
 $NP(DN@DN) \rightarrow NNP(DN) NN(DN)$





4. THỰC NGHIỆM

4.2. Xử lý câu truy vấn

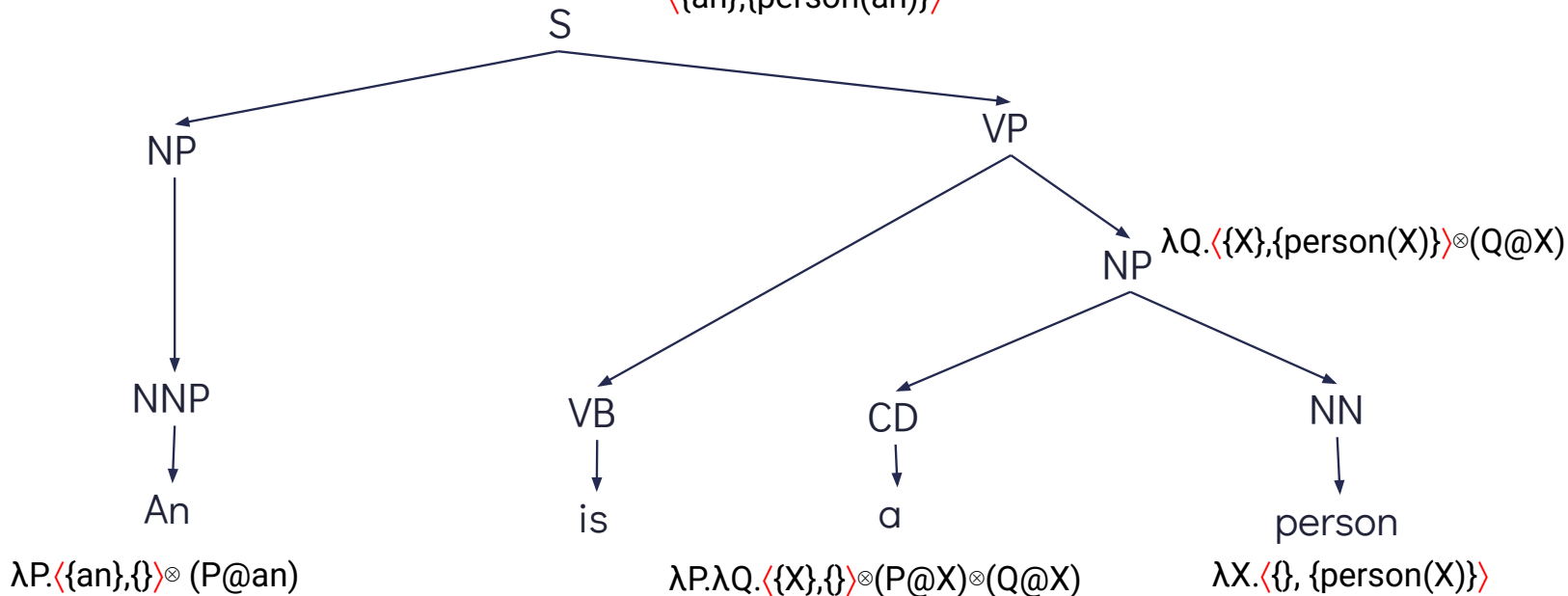


Biến đổi câu truy vấn sang FOL bằng DRS

"An is a person. He adore Vung Tau beach?"

Các bước từ 6 (Xác định biểu thức lambda) đến 8 (DRS của mỗi câu)

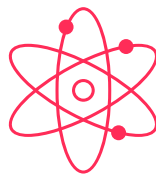
$\langle \{an\}, \{person(an)\} \rangle$





4. THỰC NGHIỆM

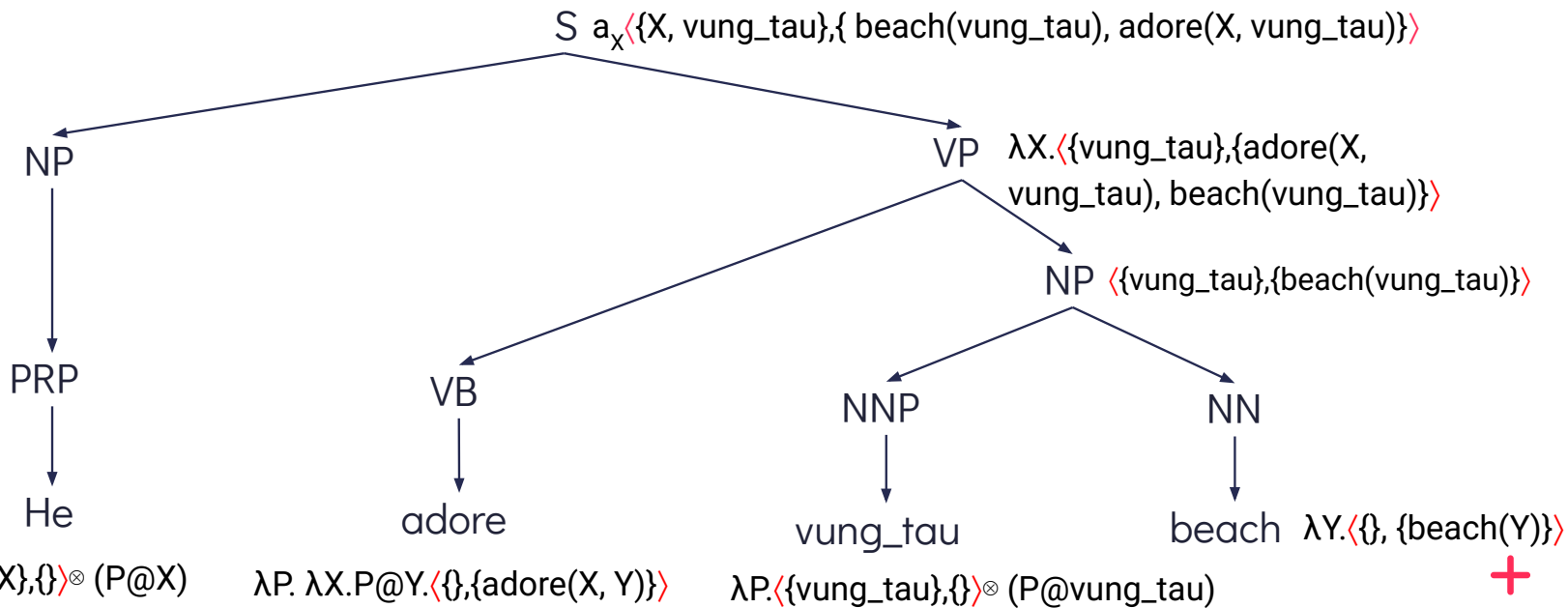
4.2. Xử lý câu truy vấn



Biến đổi câu truy vấn sang FOL bằng DRS

“An is a person. He adore Vung Tau beach?”

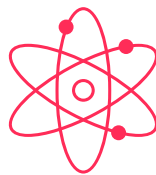
Các bước từ 6 (Xác định biểu thức lambda) đến 8 (DRS của mỗi câu)





4. THỰC NGHIỆM

4.2. Xử lý câu truy vấn



Biến đổi câu truy vấn sang FOL bằng DRS

“An is a person. He adore Vung Tau beach?”

8. Xác định DRS của mỗi câu

- DRS câu 1: $\langle \{an\}, \{person(an)\} \rangle$
- DRS câu 2: $\langle \{X, vung_tau\}, \{beach(vung_tau), adore(X, vung_tau)\} \rangle$

9. Xác định DRS của văn bản

$\langle \{an, X, vung_tau\},$
 $\{$
 $person(an),$
 $beach(vung_tau),$
 $adore(X, vung_tau),$
 $an=X$
 $\} \rangle$

10. Dịch DRS của văn bản sang FOL

$\exists x \exists y (x = an \wedge y = vung_tau \wedge person(x) \wedge beach(y) \wedge adore(x, y))$



Biến đổi câu truy vấn sang Prolog



Câu hỏi ngôn ngữ tự nhiên	Prolog query	Kết quả	Suy luận
Is Nguyen from Vung Tau?	from(nguyen, vt).	FALSE	Dựa trên FACTS: from(nguyen, gl).
Who loves Vung Tau's beach?	love(X, bh).	X= ld, lm, nd, ngan, nguyen, nm, td, thao, tm.	Rule: love(X, bh) :- person(X). person(ld), person(lm), person(nd), person(nm), person(td), person(tm), person(ngan), person(thao), person(nguyen),



Biến đổi câu truy vấn sang Prolog

Câu hỏi ngôn ngữ tự nhiên	Prolog query	Kết quả	Suy luận
Are Ngan and Thao school friends?	schoolfriend(ngan, thao).	TRUE	FACT: schoolfriend(ngan, thao).
Are Ngan and Thao pupil at Bach Khoa university?	pupil(ngan), pupil(thao), at(ngan, bk), at(thao, bk).	FALSE	FACT: pupil(ngan), pupil(thao), at(ngan, uit), at(thao, uit).



Biến đổi câu truy vấn sang Prolog

Câu hỏi ngôn ngữ tự nhiên	Prolog query	Kết quả	Suy luận
Vung Tau is a coastal city. Is beach in that beautiful and refreshing?	coastal(vt), city(vt), beach(bh) in(bh, vt) beautiful(bh), refreshing(bh)	TRUE	Fact: coastal(vt), city(vt), in(bh, vt) beautiful(bh), refreshing(bh).
Who is a student at UIT?	student(X), at(X, uit).	X = ngan, thao, nguyen	FACTS cơ sở: student(thao), student(ngan), at(thao, uit), at(ngan,uit).



Câu hỏi ngôn ngữ tự nhiên	Prolog query	Kết quả	Suy luận
What is Nguyen's mom job?	s(nguyen, lm), job(lm, X).	X= agriculturist; X = beekeeper; X = contadino; X = creator; X = dairyman; X = farmer; X = forester; X = granger; X = husbandman; X = planter; X = rancher; X = small_farmer; X = smallholder; X = sodbuster; X = sower; X = stockman; X = tenant_farmer; X = tiller(farmer);	FACT: s(nguyen, lm), job(lm, agriculturist), job(lm, beekeeper), job(lm, contadino), job(lm, creator), job(lm, dairyman), job(lm, farmer), job(lm, forester), job(lm, granger), job(lm, husbandman), job(lm, planter), job(lm, rancher), job(lm, small_farmer), job(lm, smallholder), job(lm, sodbuster), job(lm, sower), job(lm stockman,), job(lm, tenant_farmer), job(lm,), job(lm,tiller(farmer))



Biến đổi câu truy vấn sang Prolog



Câu hỏi ngôn ngữ tự nhiên	Prolog query	Kết quả	Suy luận
An is a person. Does he adore Vung Tau beach?	adore(an, bh).	FALSE	Rule: adore(X, bh) :- person(X).
Thao has a father and a mother. Is her father a fisherman and her mother a housewife?	s(thao, tm), s(thao, td), fisherman(td), housewife(tm).	TRUE	Rule: s(thao, tm), s(thao, td)fisherman(td), housewife(tm).

ỨNG DỤNG



Prolog Query App

Ứng dụng cho phép bạn query knowledge base bằng **Prolog**.



Chú thích các ký hiệu

ngan: Ngan | **nguyen:** Nguyen | **thao:** Thao | **uit:** University of Information Technology | **gl:** Gia Lai | **vt:** Vung Tau | **bh:** Beach | **td:** Thao's dad | **tm:** Thao's mom | **nd:** Ngan's dad | **nm:** Ngan's mom | **ld:** Nguyen's dad | **lm:** Nguyen's mom



Xem facts ngắn gọn

Load facts ngắn gọn



Xem tất cả facts trong Knowledge Base

Con người

ld, lm, nd, ngan, nguyen, nm, td, thao, tm

Sinh viên ở UIT

ngan, nguyen, thao

Bạn bè

ngan-nguyen, ngan-thao, nguyen-ngan, nguyen-thao, thao-ngan, thao-nguyen

Thành phố

gl, vt

Nghề nghiệp

ld-agriculturist, lm-agriculturist, td-angler, ld-beekeeper, lm-beekeeper, ld-contadino, lm-contadino, ld-creator, lm-creator, nd-creator, ngan-creator, nguyen-creator, nm-creator, td-creator,

ỨNG DỤNG

Prolog Query App


Ứng dụng cho phép bạn query knowledge base bằng Prolog.

Chú thích các ký hiệu

ngan: Ngan | **nguyen:** Nguyen | **thao:** Thao | **uit:** University of Information Technology | **gl:** Gia Lai | **vt:** Vung Tau | **bh:** Beach | **td:** Thao's dad | **tm:** Thao's mom | **nd:** Ngan's dad | **nm:** Ngan's mom | **ld:** Nguyen's dad | **lm:** Nguyen's mom

▼  Xem facts ngắn gọn

Load facts ngắn gọn

>  Xem tất cả facts trong Knowledge Base



```
quitter/1
scratcher/1
someone/1
tonic/1
walk_in/1
term_expansion/2
educatee/1
free_agent/1
immune/1
alter_ego/2
juvenile/1
lover/1
mortal/1
ouster/1
primitive/1
auditor/1
buster/1
contadino/1
town/2
scorpio/1
radical/1
tiger/1
```

ỨNG DỤNG



Thực hiện query Prolog

Nhập câu lệnh Prolog:

```
pupil(ngan), pupil(thao), at(ngan, bk), at(thao, bk).
```

Run Query

☒ Result: False

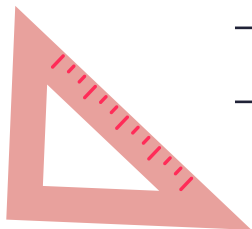
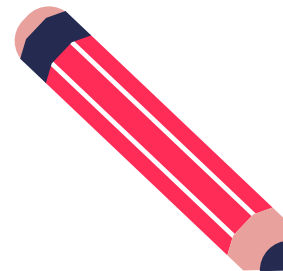


[CS229WordnetlabelDemo](#)

NHẬN XÉT & KẾT LUẬN

- Nhận xét
- Hạn chế
- Kết luận

05





NHẬN XÉT

ƯU ĐIỂM



Truy vấn chính xác những thông tin cơ bản (vị trí, đối tượng,...) trong fact và rule.

What is Ngan's hometown?

Who are students at UIT?

Tạo logic giữa các thực thể khác nhau trong văn bản thông qua các khái niệm chung trong bộ từ điển.





NHẬN XÉT

ƯU ĐIỂM



Trả lời được các câu hỏi sử dụng từ vựng khác với văn bản gốc, có thể hỏi 1 câu cùng ý nghĩa bằng nhiều cách.

Giúp hệ thống hiểu rằng fisherman, fisher, trawler đều hướng về một khái niệm.

Is Thao's dad a fisherman/ fisher / trawler?



Cung cấp quan hệ Hypernym và Hyponym cho phép Prolog suy luận bắc cầu.

Ví dụ hypernym của các nghề nghiệp như fisherman, housewife, fishmonger là job, có thể tạo rule:
job(X, fisher):- person(X), fisherman(X);
trawler(X).



HẠN CHẾ

Nhiều Nghĩa từ WordNet

Không sát ngữ cảnh, Tri thức bị "loãng" do gán cả những nghĩa hiếm hoặc nghĩa ở chuyên ngành khác vào một văn bản đời thường.

→ cơ sở tri thức bị dàn trải và chứa nhiều thông tin sai ngữ cảnh.

Khi truy vấn nghề nghiệp của mẹ Nguyễn (farmer), hệ thống trả về cả beekeeper, sodbuster, creator

HẠN CHẾ

Tính Đa hình trong Prolog

Prolog khó xử lý nhất quán các vị từ (predicates) có số lượng tham số thay đổi (arity)

→ khiến việc viết các quy tắc suy luận tổng quát trở nên phức tạp và dễ gây lỗi.

Khi định nghĩa quan hệ cha con, có thể dùng:

- **dad(thao_dad)** (đối tượng này là một người cha)
- **dad(thao_dad, thao)** (người này là cha của Thảo).

Trong Prolog, đây là hai vị từ hoàn toàn khác nhau (dad/1 và dad/2)

HẠN CHẾ

Phụ thuộc vào Thế giới đóng

Hệ thống chỉ có thể suy luận trên các thực thể đã khai báo sẵn

→ Các thực thể mới nằm ngoài dữ liệu đầu vào sẽ bị trả về kết quả Sai (False) dù đúng ngữ cảnh.

∀ y person(y) → loves(y, bh).
Nếu trong hệ thống chỉ khai báo person(ngan), person(thao), thì khi hỏi "An có yêu biển không?"

→ False vì nhân vật "An" chưa được định nghĩa là một person trong KB, dù thực tế "An" là con người.

HẠN CHẾ

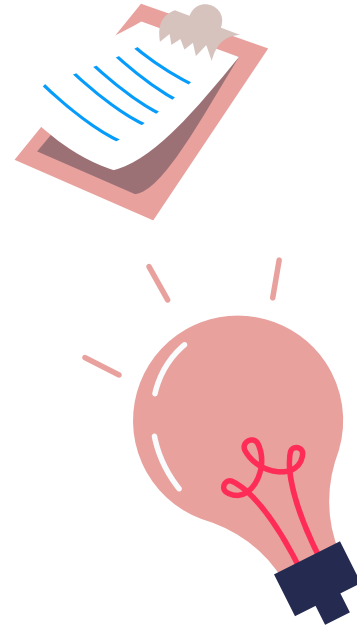
Đòi hỏi có kiến thức về Prolog

→ Chưa hỗ trợ hoàn toàn việc nhập liệu bằng ngôn ngữ tự nhiên phức tạp.

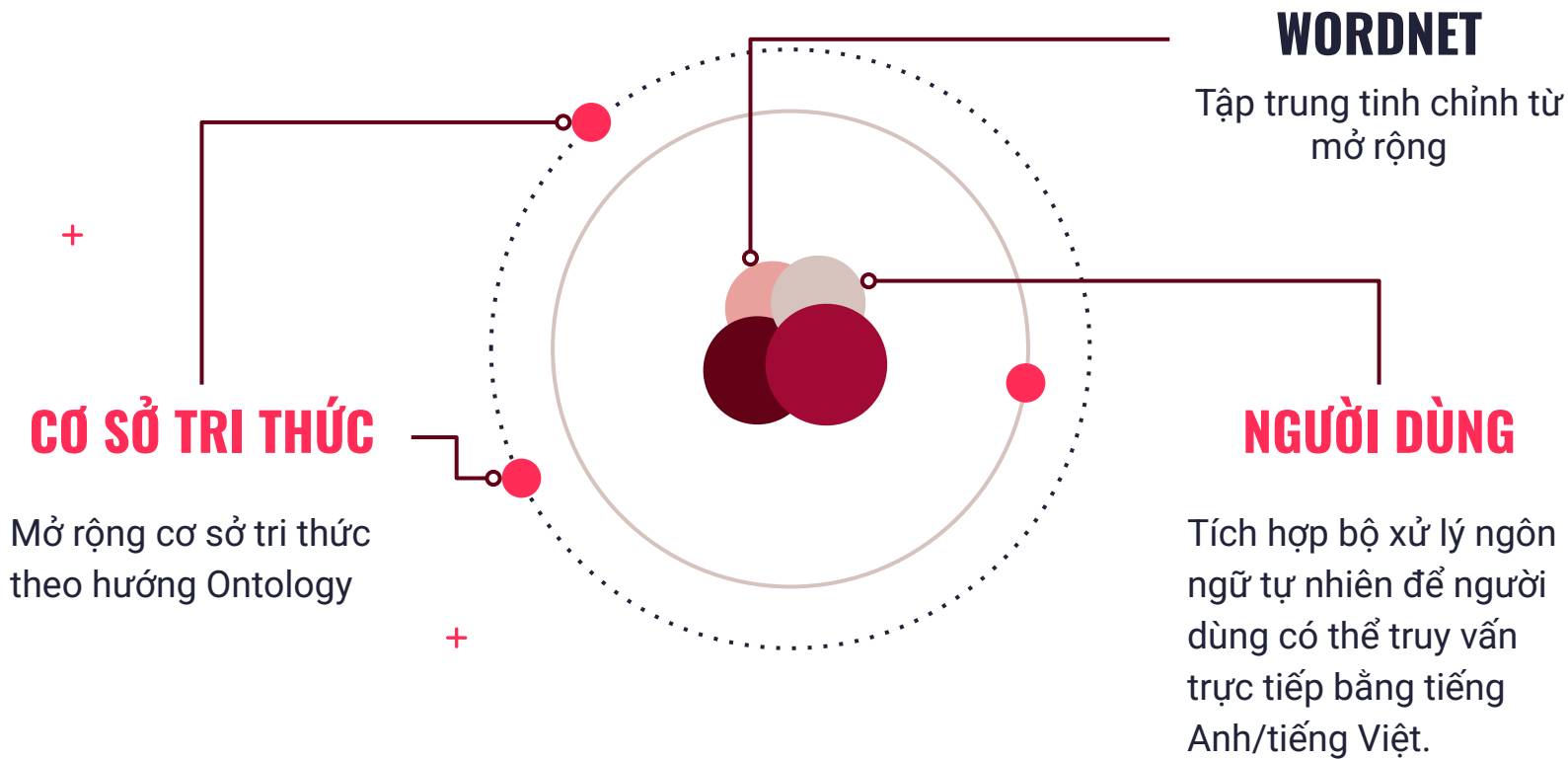
Muốn biết "Ai học ở UIT?", người dùng vẫn phải nhập đúng cú pháp Prolog là `student(X), at(X, uit).`

KẾT LUẬN

- ☒ Triển khai thành công quy trình từ xử lý ngữ nghĩa (WSD)
- ☐ Hỗ trợ người dùng phổ thông.
- ☒ Xây dựng ứng dụng web truy vấn tri thức trực quan.
- ☒ Mô phỏng chính xác quá trình suy luận dựa trên các quy tắc (Rules) và dữ kiện (Facts) trong Prolog.
- ☐ Cân bằng giữa việc mở rộng tri thức (để hiểu nhiều hơn) và kiểm soát nhiễu (để suy luận đúng hơn).
- ☒ Biểu diễn logic bậc một (FOL)



HƯỚNG PHÁT TRIỂN



HƯỚNG PHÁT TRIỂN

CƠ SỞ TRI THỨC

Sử dụng Arity cao nhất

Luôn sử dụng vị từ có 2 tham số để mô tả quan hệ. Nếu không biết con là ai, có thể dùng biến vô danh (_).

- `dad(thao_dad, thao).`
- `dad(nguyen_dad, _).` (Chỉ biết là cha, chưa rõ con là ai)

Tách biệt Loại thực thể và Quan hệ

Sử dụng các tên vị từ khác nhau hoàn toàn để tránh nhầm lẫn:

- `is_father(thao_dad).` (Vị từ 1 ngôi - chỉ loại thực thể/ngành nghề)
- `father_of(thao_dad, thao).` (Vị từ 2 ngôi - chỉ mối quan hệ)

HƯỚNG PHÁT TRIỂN

WORDNET

Giải pháp Khử nhập nhằng (WSD Strategies): Để khắc phục tình trạng nhiễu tri thức, hệ thống cần áp dụng các kỹ thuật khử nhập nhằng nâng cao:

- Sử dụng Ngữ cảnh Cục bộ (Local Context): Chỉ chọn những từ đồng nghĩa có xuất hiện trong cửa sổ văn bản xung quanh.
- Áp dụng Lesk Algorithm: Tính toán mức độ trùng lặp từ vựng giữa định nghĩa của các nghĩa từ với các từ xung quanh.
- Mô hình Ngôn ngữ Pre-trained (RoBERTa/BERT): Tận dụng embeddings để xác định Synset gần nhất với ngữ cảnh hiện tại.
- Hạn chế bậc mở rộng: Giới hạn khoảng cách trong cây phân cấp WordNet để tránh việc suy luận quá xa rời thực tế.

**CẢM ƠN THẦY
VÀ CÁC BẠN ĐÃ
LẮNG NGHE!**

