



Bach Khoa
University

Start



KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

**MỘT ĐỀ XUẤT KHÁM PHÁ THỰC THỂ DỰA TRÊN Ý ĐỊNH
ĐỂ XÂY DỰNG ĐỒ THỊ TRI THỨC TRONG LĨNH VỰC GIÁO
DỤC TẠI TRƯỜNG ĐẠI HỌC BÁCH KHOA – ĐHQG-HCM**

GVHD: PGS. TS. Bùi Hoài Thắng
ThS. Bùi Công Tuấn

SVTH: Võ Ngọc Duy Nghiêm
Nguyễn Minh Khoẻ



NỘI DUNG

01 GIỚI THIỆU ĐỀ TÀI

**02 CƠ SỞ LÝ THUYẾT
CÔNG TRÌNH LIÊN QUAN**

03 PHƯƠNG PHÁP

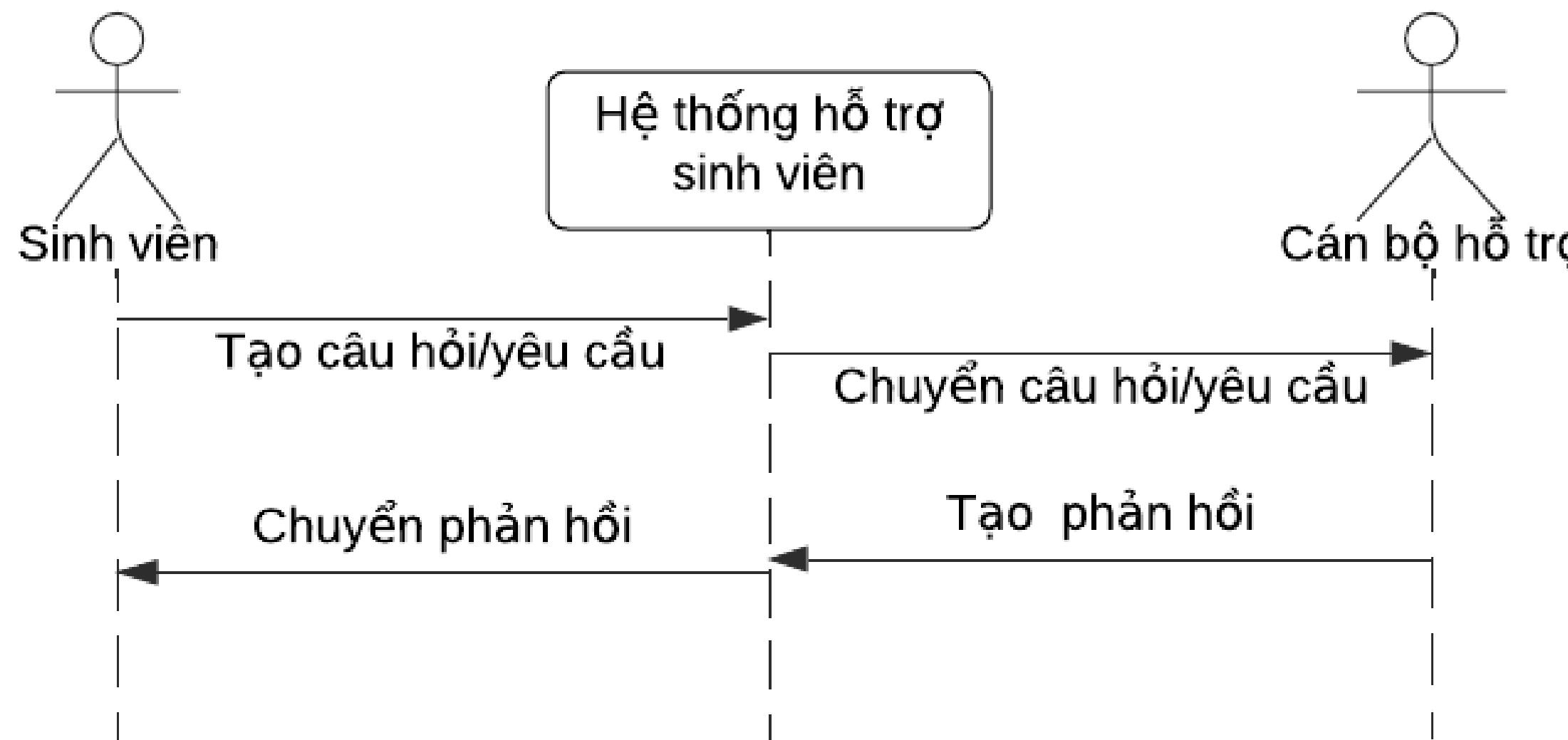
04 THÍ NGHIỆM

05 TỔNG KẾT

1. Giới thiệu đề tài

02

1.1 Đặt vấn đề



Hình 1. Sơ đồ luồng xử lý của hệ thống hỗ trợ sinh viên BKSI



1.1 Đặt vấn đề

STT	Chủ đề	Số lượng câu hỏi/yêu cầu
1	Đăng ký môn học	8.288
2	Tư vấn học vụ	4.276
3	Tốt nghiệp - Điểm - Lịch thi	6.875
4	Yêu cầu (nộp/khiếu nại)	4.821
5	Học bổng	229
6	Sinh hoạt công dân	320
7	Miễn giảm và gia hạn học phí	91
8	Y tế	286
9	Xác nhận sinh viên vay vốn	73
10	Chuẩn ngoại ngữ	1.200
11	Tiếng nhật	5
12	Học phí	718
13	Khác	450
	Tổng cộng	27.632

Bảng 1: Thống kê số lượng câu hỏi theo chủ đề của BKSI trong 2 năm (2021-2022)



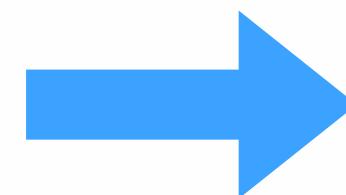
1.1 Đặt vấn đề

Câu hỏi/yêu cầu có thể dài dòng, chứa nhiều thông tin lan man

Kính thưa PĐT,
Em tên là Nguyễn Văn A, học lớp MTKH01, khoa
Khoa học và Kỹ thuật Máy tính. Vào ngày hôm qua
(10/1/2019), em có đăng ký môn học Anh văn 4
nhưng không thành công vì sĩ số đã đầy.

Em viết đơn này để xin phòng đào tạo có thể cân
nhắc tăng sĩ số lớp A01 môn Anh văn 4 (mã môn
học là IA1009) vào kì dự thi 2019-2020

Em xin cảm ơn!



Ý định: tăng sĩ số lớp
Môn học: Anh văn 4
Mã môn học: IA1009
Nhóm lớp: A01
Học kì: dự thi 2019 - 2020



1.1 Đặt vấn đề

Đồ thị tri thức

- Giúp nhận dạng, phân tích các ý định hỏi và thông tin liên quan của sinh viên.
- Giúp trả lời các câu hỏi một cách nhanh chóng và chính xác.



1.2 Mục tiêu đề tài

Mục tiêu 1: Đề xuất framework để xây dựng đồ thị tri thức tập trung vào ý định.

Mục tiêu 2: Đề xuất framework cho nhiệm vụ phát hiện ý định và điền trường thông tin.



1.3 Nhiệm vụ trong đề tài

Đề tài bao gồm 2 pha:

Pha 1: Xây dựng đồ thị tri thức

- Đầu vào: bộ dữ liệu gồm lịch sử các câu hỏi của sinh viên.
- Đầu ra: các bộ ba cho đồ thị tri thức.

Pha 2: Phát hiện ý định và điền trường thông tin

- Đầu vào: câu hỏi mới
- Đầu ra: bộ ba mới để tích hợp thêm vào đồ thị tri thức.



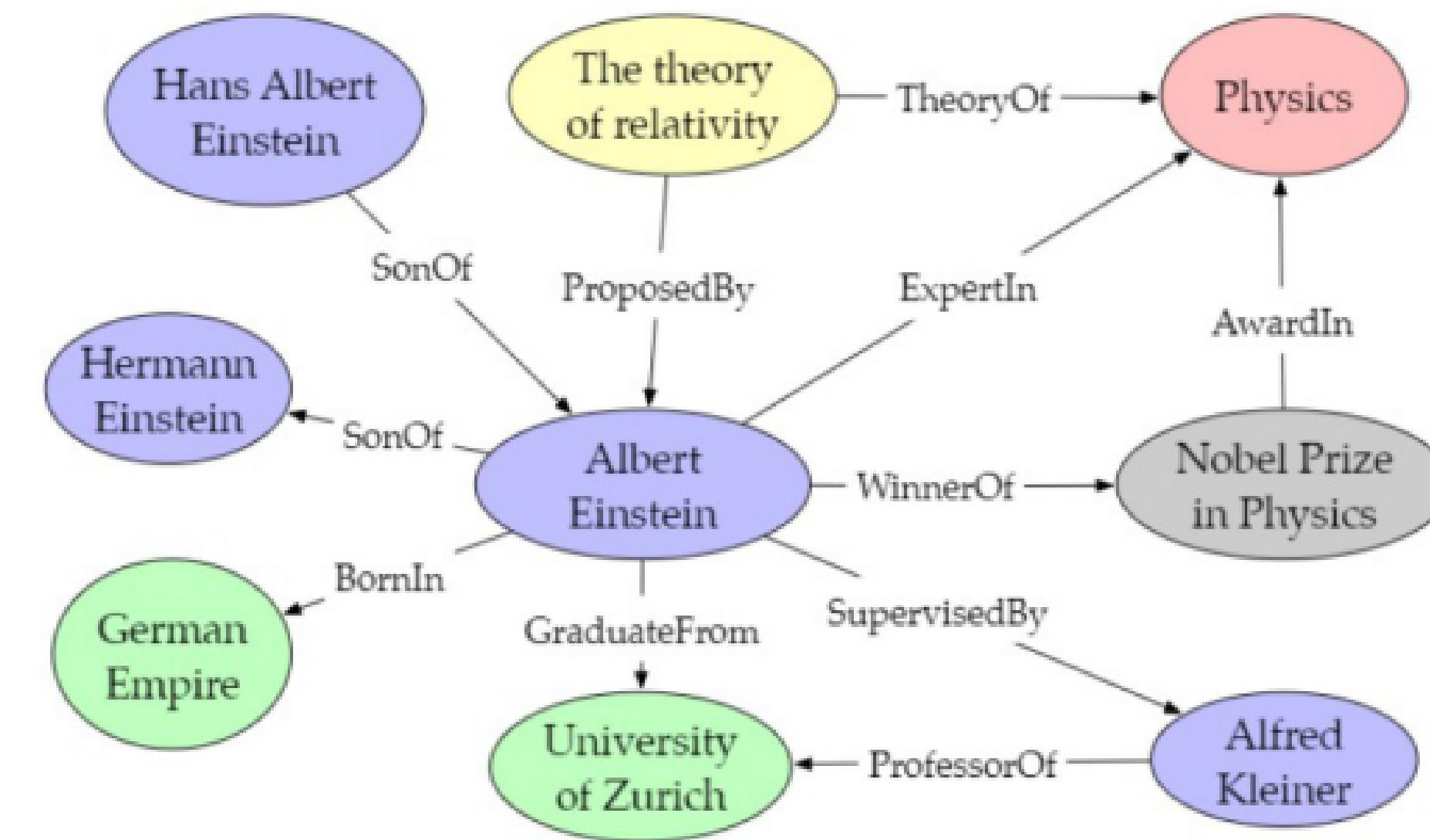
1.4 Phạm vi đề tài

- **Dữ liệu:** Các câu hỏi gửi đến BKSI của Phòng Đào tạo và các nguồn website của trường Đại học Bách Khoa - ĐHQG TP. Hồ Chí Minh.
- **Ứng dụng:** chỉ giúp giải quyết vấn đề tóm gọn thông tin trong câu hỏi, hỗ trợ chuyên viên trong việc xác định nội dung câu hỏi một cách nhanh chóng.

2. Cơ sở lý thuyết và Công trình liên quan

2.1 Đồ thị tri thức

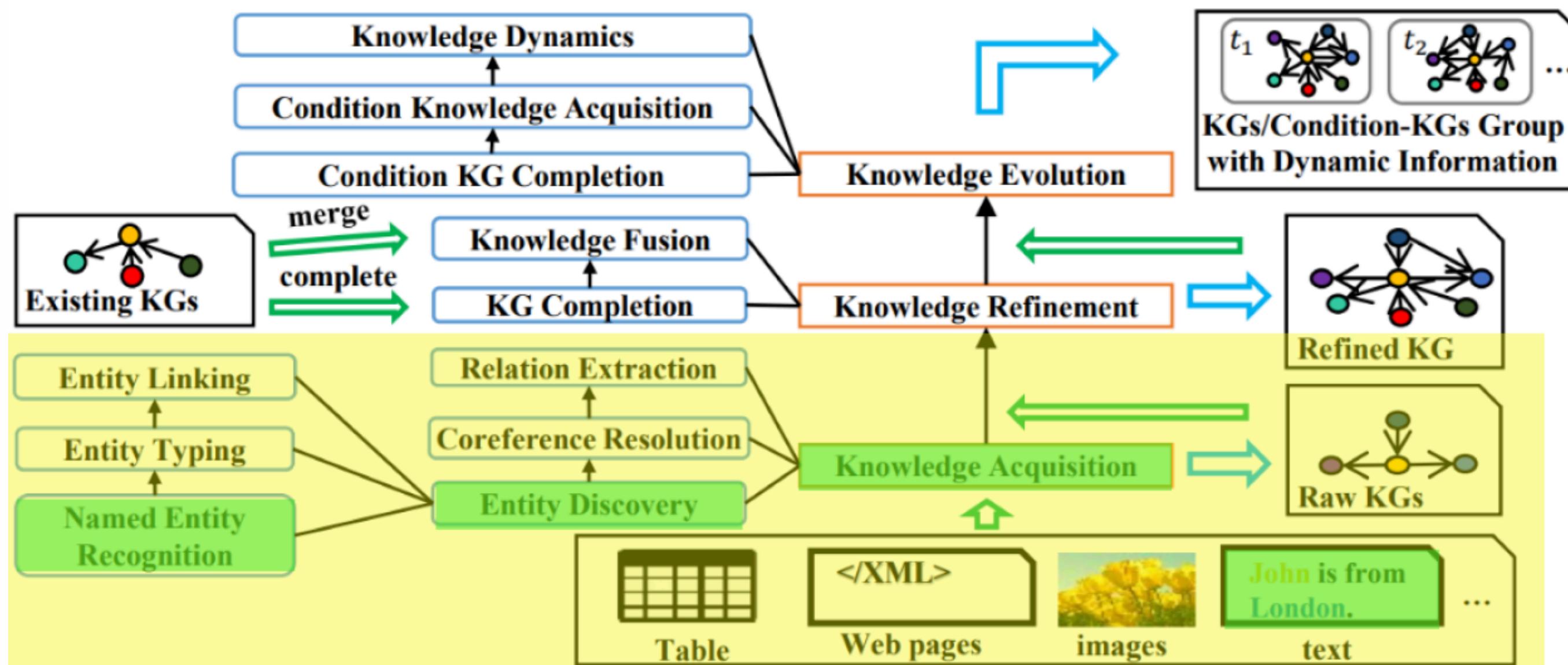
- Thực thể
- Kiểu thực thể
- Mối quan hệ
- Lược đồ



Hình 2. Đồ thị tri thức [2]

2. Cơ sở lý thuyết và Công trình liên quan

2.1 Đồ thị tri thức



Hình 3. Quy trình xây dựng đồ thị tri thức [6]



2. Cơ sở lý thuyết và Công trình liên quan

2.1 Đồ thị tri thức

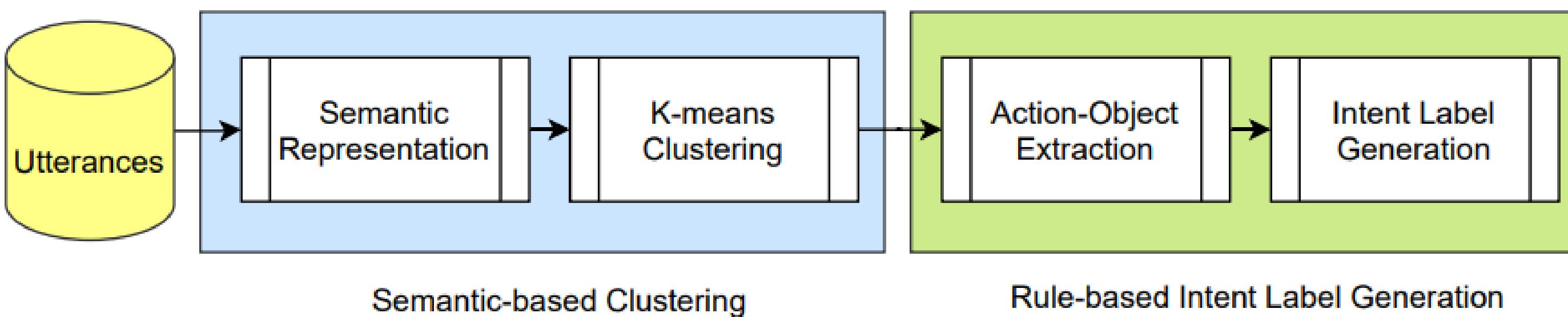
Hướng tiếp cận khi xây dựng đồ thị tri thức:

- Top - down: rút trích thông tin dựa trên lược đồ được định nghĩa trước.
- Bottom - up: rút trích thông tin trước, xây dựng lược đồ từ tri thức.

2. Cơ sở lý thuyết và Công trình liên quan

2.2 Trích xuất thông tin mở

Trích xuất thông tin mở (Open Information Extraction) là một nhiệm vụ trong xử lý ngôn ngữ tự nhiên nhằm mục đích tự động trích xuất các thông tin quan trọng từ văn bản mà không cần biết trước các kiểu thực thể hoặc quan hệ cụ thể.



Hình 4. Khám phá ý định mở [1]



2.3 Nhận diện thực thể có tên (NER)

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

A little less than a decade later DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Hình 5. Minh họa cho nhiệm vụ nhận diện thực thể có tên



2.4 Phân loại ý định và điền trường thông tin

Phân loại ý định là nhiệm vụ phát hiện mục đích của một câu truy vấn.

Điền trường thông tin là tìm kiếm các thông tin làm rõ ràng hơn ý định để hỗ trợ xử lý một truy vấn

query	find	recent	comedies	by	james	cameron
slots	O	B-date	B-genre	O	B-dir	I-dir
intent	find_movie					

Bảng 2: Minh họa cho nhiệm vụ phân loại ý định và điền trường thực thể



2. Cơ sở lý thuyết và Công trình liên quan

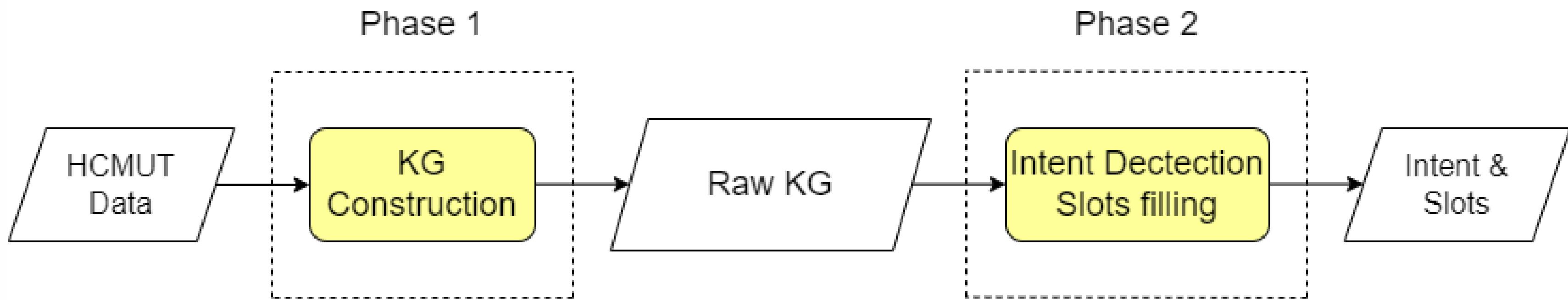
2.5 Đọc hiểu máy (MRC)

MRC là nhiệm vụ giúp máy có thể đọc hiểu một văn bản tự nhiên và trả lời các câu hỏi liên quan đến nội dung trong văn bản.

$$a = f(c, q)$$

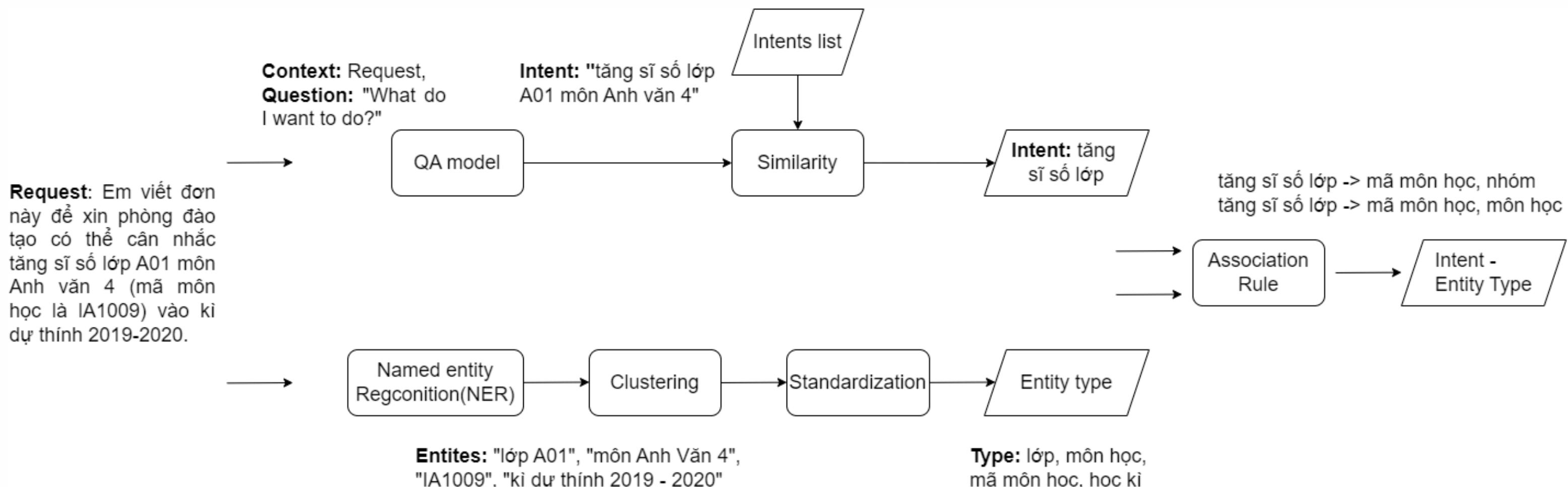
- c: văn bản dùng để trả lời câu hỏi (ngữ cảnh)
- q: câu hỏi hỏi về nội dung ngữ cảnh
- f: mô hình đọc hiểu máy
- a: câu trả lời

3. Phương pháp



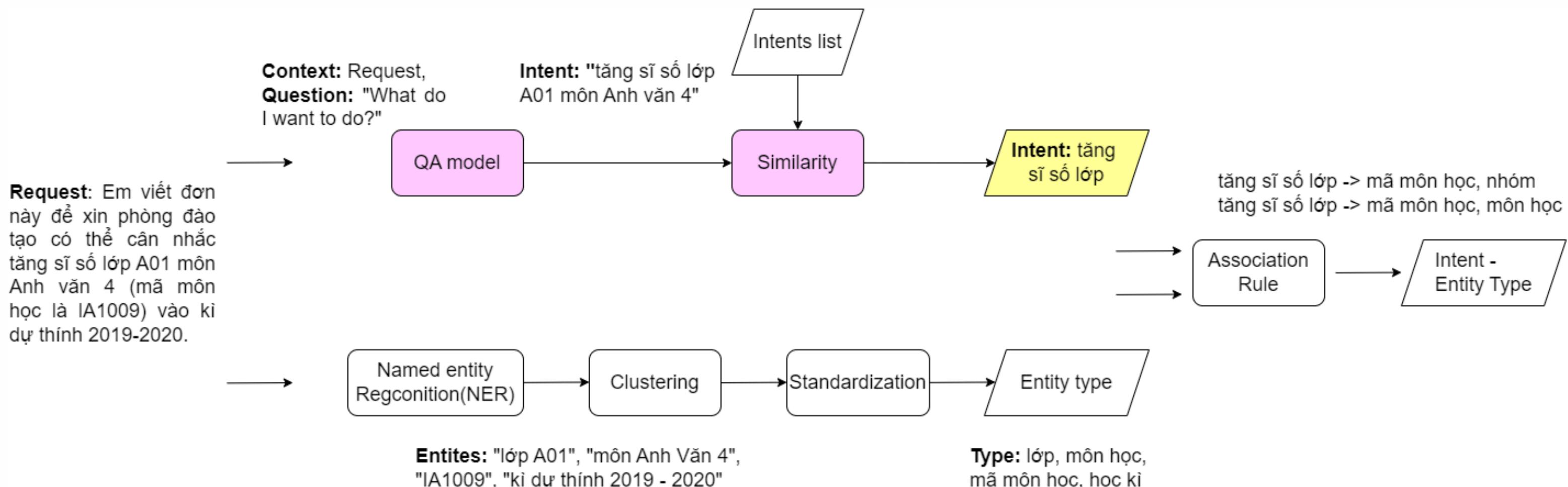
Hình 6: Framework tổng cho bài toán

3.1 Xây dựng đồ thị tri thức dựa trên ý định



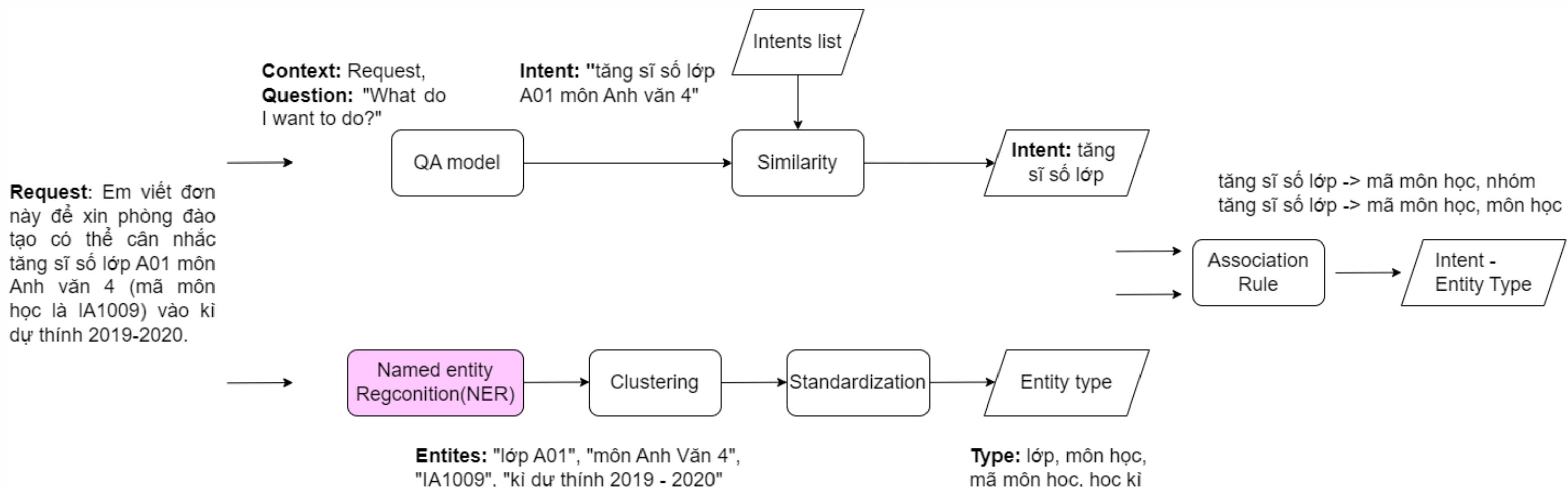
Hình 7: Framework NCA cho nhiệm vụ xây dựng đồ thị tri thức xoay quanh ý định

3.1 Xây dựng đồ thị tri thức dựa trên ý định



Hình 7: Framework NCA cho nhiệm vụ xây dựng đồ thị tri thức xoay quanh ý định

3.1 Xây dựng đồ thị tri thức dựa trên ý định



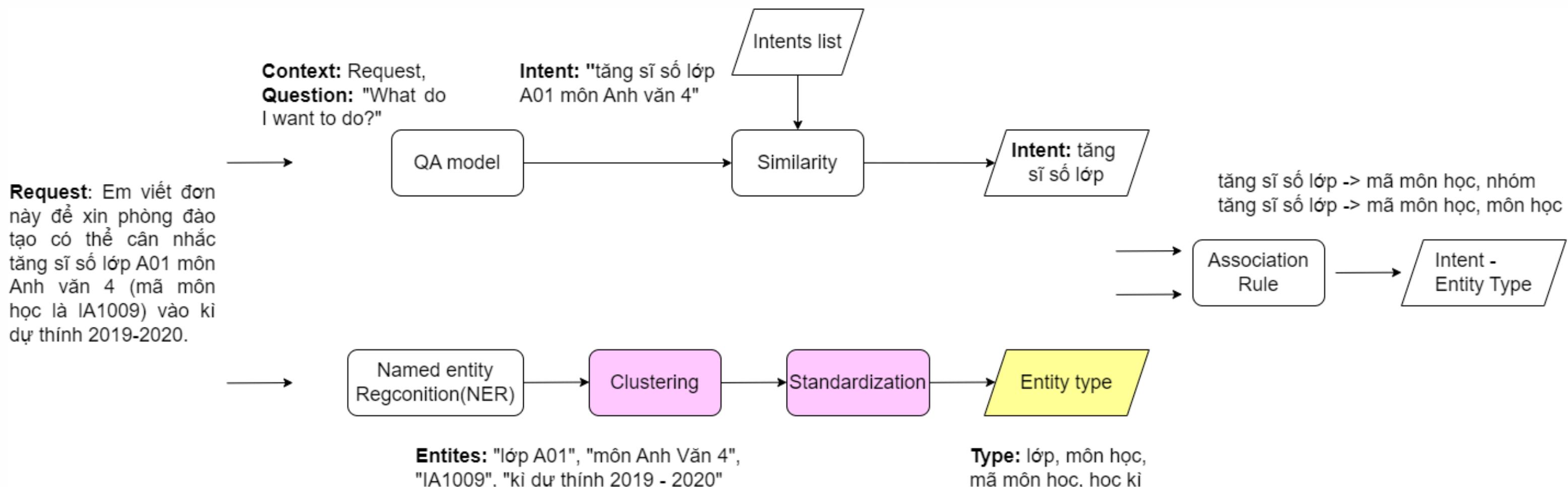
Hình 7: Framework NCA cho nhiệm vụ xây dựng đồ thị tri thức xoay quanh ý định

3.1 Xây dựng đồ thị tri thức dựa trên ý định

	VLSP	Nhóm
PER	Tên của người, động vật hay nhân vật hư cấu và danh xưng	Giống VLSP
ORG	Tên các tổ chức, cơ quan, công ty, thương hiệu	Giống VLSP
LOC	Các thực thể có tọa độ địa lý nhất định	Giống VLSP
MISC	Dùng để giải quyết một số trường hợp nhập nhằng và markup tên tác phẩm, sự kiện, thương hiệu có chứa tên thuộc 3 loại trên	Lợi dụng nhãn này để gán cho những thực thể chưa rõ kiểu có trong miền giáo dục

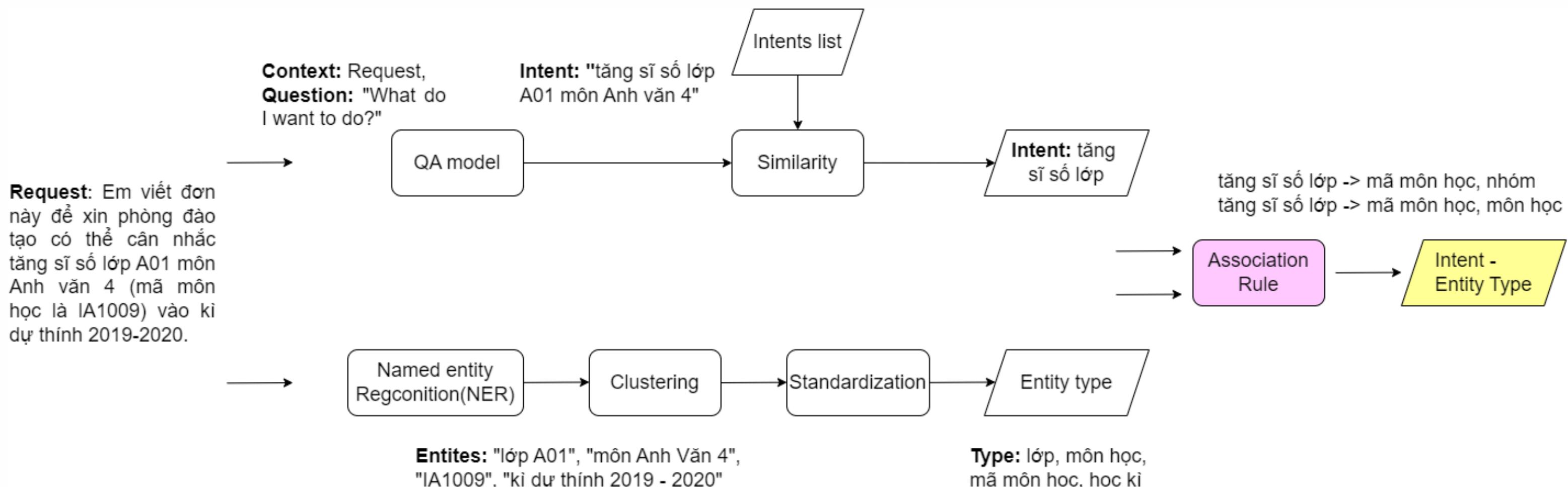
Bảng 3: So sánh cách đánh nhãn của VLSP 2018 và nhóm cho nhiệm vụ NER

3.1 Xây dựng đồ thị tri thức dựa trên ý định



Hình 7: Framework NCA cho nhiệm vụ xây dựng đồ thị tri thức xoay quanh ý định

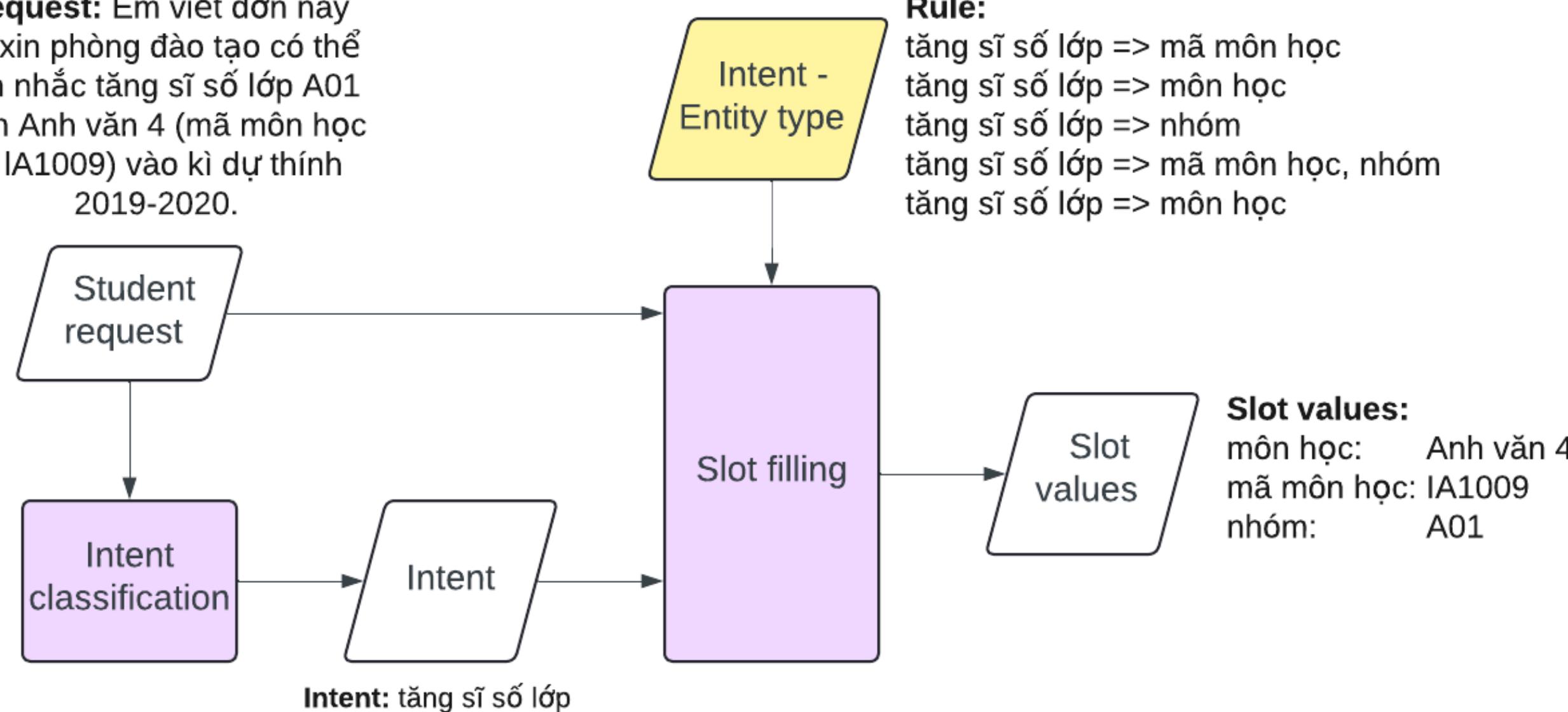
3.1 Xây dựng đồ thị tri thức dựa trên ý định



Hình 7: Framework NCA cho nhiệm vụ xây dựng đồ thị tri thức xoay quanh ý định

3.2 Phân loại ý định & Điền trường thông tin

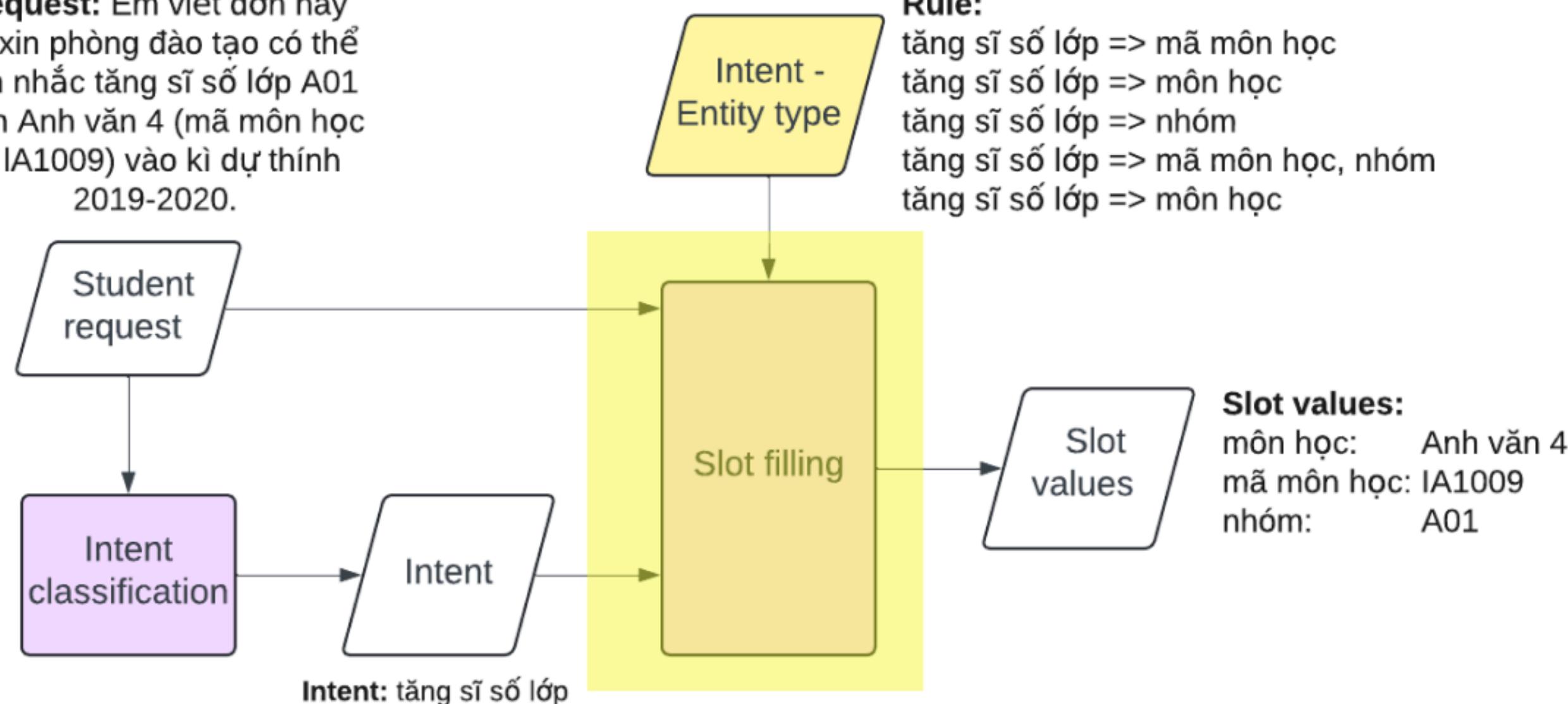
Request: Em viết đơn này để xin phòng đào tạo có thể cân nhắc tăng sĩ số lớp A01 môn Anh văn 4 (mã môn học là IA1009) vào kì dự thính 2019-2020.



Hình 8 : Framework cho Phân loại ý định & Điền trường thông tin

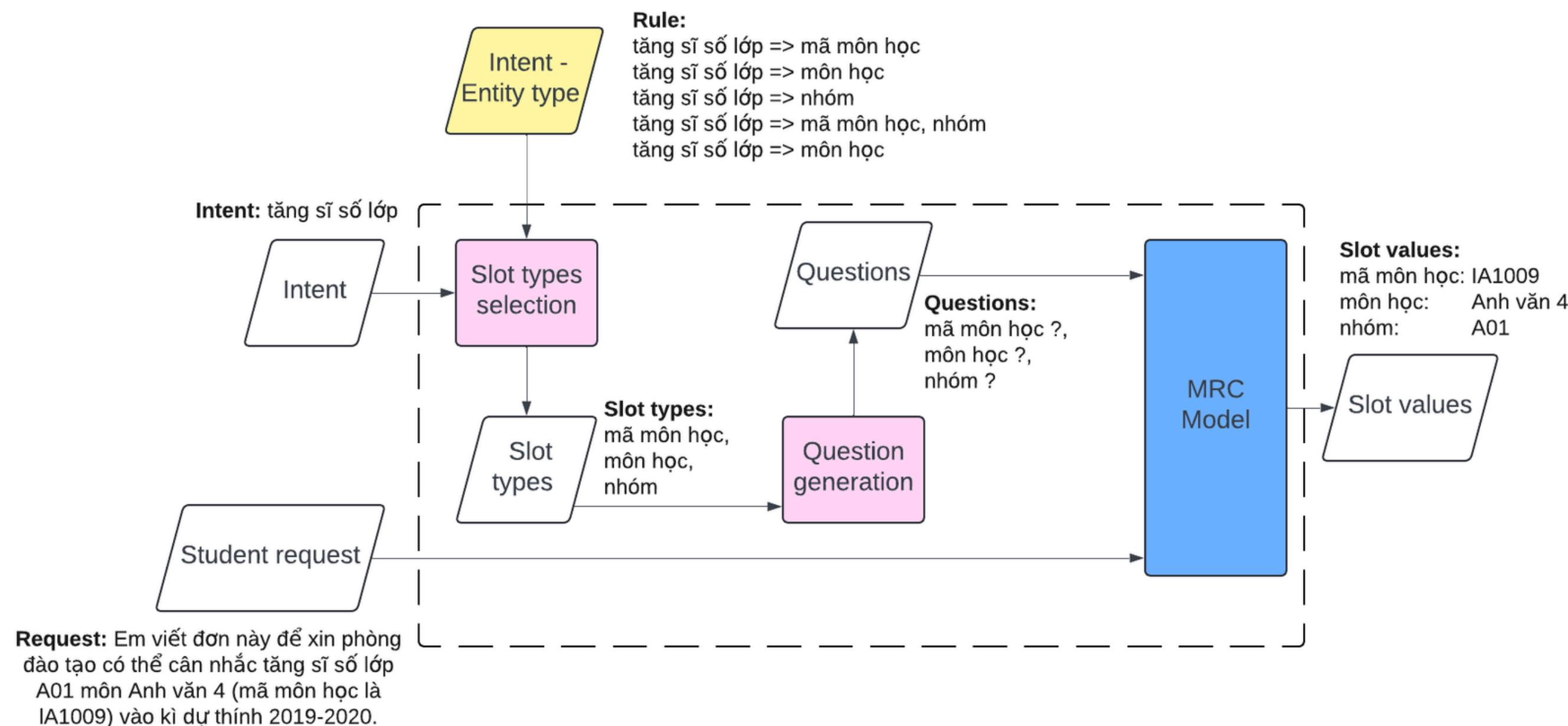
3.2 Phân loại ý định & Điền trường thông tin

Request: Em viết đơn này để xin phòng đào tạo có thể cân nhắc tăng sĩ số lớp A01 môn Anh văn 4 (mã môn học là IA1009) vào kì dự thính 2019-2020.



Hình 8 : Framework cho Phân loại ý định & Điền trường thông tin

3.2 Phân loại ý định & Điền trường thông tin



Hình 9 : Framework cho Điền trường thông tin



3.2 Điền trường thông tin

Câu hỏi được tạo ra dựa trên chiến lược sinh theo luật (rule-based)

Các mẫu câu hỏi do nhóm tự đề xuất

Mẫu 1 : kiểu thực thể + “?”

Mẫu 2 : kiểu thực thể + “là gì” + “?”

Mẫu 3 : “Sinh viên muốn” + ý định + “với” + kiểu thực thể + “là gì” + “?”



3.3 Cấu trúc bộ ba đồ thị tri thức

Sử dụng cấu trúc bộ ba hyper relational triplet

(Sinh viên, ý định, thực thể 1)

- attribute: thực thể 2, thực thể 3, ...

Request	Em viết đơn này để xin phòng đào tạo có thể cân nhắc tăng sĩ số lớp A01 môn Anh văn 4 (mã môn học là IA1009) vào kì dự định 2019-2020
Hyper relational triplet	(N/A, [tăng sĩ số lớp, {mã môn học: IA1009, nhóm: A01}], Anh văn 4)



4.1 Nhận diện thực thể có tên

Dữ liệu gán nhãn:

- Raw_Data: Gồm các thông tin từ các website chính thức của nhà trường
- FAQs: Hơn 200,000 câu hỏi/yêu cầu đến từ sinh viên được lấy từ hệ thống

BKSI: **1%** đánh nhãn

Dữ liệu để trích thực thể: **10%** FAQs

Mô hình [NlpHUST/her-vietnamese-electra-base](#)

- ELECTRA [5]
- Dữ liệu huấn luyện: VLSP NER 2018.

4.1 Nhận diện thực thể

Class	P	R	F1
PER	0,00	0,00	0,00
ORG	33,33	53,33	41,03
LOC	42,86	50,00	46,15
MISC	20,00	0,44	0,85
Overall	30,23	9,22	14,13

Bảng 4: Kết quả đánh giá mô hình trước khi tinh chỉnh

Class	P	R	F1
PER	100,00	50,00	66,66
ORG	37,68	57,77	45,61
LOC	42,86	50,00	46,15
MISC	68,96	69,87	69,41
Overall	61,49	67,38	64,29

Bảng 5: Kết quả đánh giá mô hình sau khi tinh chỉnh

4.2 Gom cụm

No	Embedding Model	Dimensions	Min cluster size	Num of Clusters	Silhouette Score
1	simCSE	50	30	66	0.599
2	simCSE	9	60	26	0.601
3	SBERT	16	10	231	0.617
4	simCSE	16	50	36	0.618
5	simCSE	9	50	36	0.622
6	simCSE	60	10	36	0.628
7	SBERT	9	20	117	0.636
8	simCSE	9	20	122	0.647
9	SBERT	16	20	112	0.649
10	simCSE	9	10	241	0.65

Bảng 6: Bộ thông số dùng cho quá trình phân cụm và kết quả tương ứng



4.2 Gom cụm

No	Entity type	Entity Code	Sub-Entity Type	Example
1	Tên người	PER		
2	Tổ chức	ORG	trường, phòng ban công ty	
3	Địa điểm	LOC		cơ sở 2
4	Chương trình đào tạo	EDT		chính quy vừa học vừa làm
5	Khoa	FAC		
6	Ngành	MAJ		
7	Môn học	CRN		
8	Mã môn học	CSC		CO1027
9	Thời gian	DAT	Ngày, tháng, năm	
10	Học kì	SEM		học kì 231
11	Lớp	CLA		MT22B2KH
12	Nhóm	GRO		L01
13	Khoá	COH		K20

Bảng 7: Một số kiểu thực thể được khám phá



4.3 Phân loại ý định

Nhóm đã sử dụng **79** ý định trong tổng số **117** ý định từ danh sách có sẵn.

Mô hình nhóm sử dụng cho nhiệm vụ trả lời câu hỏi là một biến thể của mô hình XLM-RoBERTa [6].

Cả quá trình phân loại ý định bằng phương pháp đọc hiểu máy kết hợp với cosine similarity, độ chính xác đạt được là **16.5%** trên tổng số 297 nhãn ý định.



4.4 Áp dụng luật kết hợp

Ý định	Kiểu thực thể
đăng kí môn học	mã môn học, môn học
đăng kí môn học	mon học, học kì
cập nhật thời khóa biểu	học kì
hỏi kết quả đăng kí môn học	đợt, môn học
hỏi điều kiện đăng kí môn học	mã môn học, môn học
khiếu nại điểm	loại điểm, môn học
khiếu nại điểm	loại điểm, thời gian
khôi phục thời khóa biểu	mon học, học kì
rút môn học	nhóm, mã môn học
nhận giấy xác nhận sinh viên	tài liệu, thời gian

Bảng 8: Kết quả một số luật được khai phá



4.5 Điền trường thông tin

Nhóm đã thí nghiệm với ba mẫu câu hỏi và so sánh kết quả dự đoán của mô hình trên ba mẫu câu hỏi này.

Mẫu câu hỏi	P	R	F1
Mẫu 1	24,03	1,67	3,13
Mẫu 2	24,81	1,73	3,23
Mẫu 3	19,69	1,35	2,53

Hình 10: Kết quả đánh giá mô hình trên một số mẫu câu hỏi



5.1 Kết quả đạt được

Mục tiêu 1:

Đã hoàn thành

- Đề xuất framework xây dựng đồ thị tri thức theo hướng khai phá các kiểu thực thể và thí điểm trên dữ liệu trường Đại học Bách Khoa
- Xây dựng các bộ ba cho đồ thị tri thức dựa trên ý định và các kiểu thực thể liên quan

Chưa hoàn thành:

- Thực hiện tinh chỉnh lại bộ luật khai phá

Mục tiêu 2:

- Đề xuất và thử nghiệm phương pháp Đọc hiểu máy cho nhiệm vụ phát hiện ý định và điền trường thông tin.



5.2 Ưu điểm

- Đề xuất một framework khám phá các kiểu thực thể trên một tập dữ liệu mới
- Ứng dụng đồ thị tri thức được khai phá để hỗ trợ cho bài toán điền trường thông tin
- Đề xuất một cấu trúc cho đồ thị tri thức có thể chứa ý định và các thông tin liên quan



5.3 Nhược điểm

- Nhiệm vụ trích xuất thực thể cần nhiều dữ liệu gán nhãn hơn để mô hình hoạt động tốt.
- Giải thuật nhúng chưa tốt khiến việc gom cụm có xu hướng gom các span gần giống nhau về mặt biểu diễn thay vì chung kiểu thực thể
- Mô hình đọc hiểu máy cho điền trường thông tin chỉ trích xuất một span duy nhất.



5.4 Hướng phát triển

- Tìm phương pháp nhúng để phân cụm phù hợp
- Tinh chỉnh, hoàn thiện các bộ ba của đồ thị tri thức
- Tinh chỉnh lại bộ luật được khai phá
- Xử lý vấn đề đa ý định đến từ sinh viên
- Tìm hiểu và thử nghiệm các mô hình đọc hiểu máy có thể trích nhiều span.
- Hiện thực một API tích hợp tác vụ điền các trường thông tin vào hệ thống hỏi đáp BKSI để hỗ trợ chuyên viên trả lời câu hỏi

Tài liệu tham khảo

- [1] Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen M. Meng. Open intent discovery through unsupervised semantic clustering and dependency parsing.
- [2] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications.
- [3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.



Tài liệu tham khảo

- [4] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embedding
- [5] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators.
- [6] Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction.

**CHÚNG EM CHÂN
THÀNH CẢM ƠN QUÝ
HỘI ĐỒNG ĐÃ CHÚ Ý
LẮNG NGHE**