Các bước đề nghị

1 NHẬN DIỆN THỰC THỂ ĐẶT TÊN (NER)

2.TIỀN XỬ LÝ POS TAGGER

3. XÂY DỰNG DANH SÁCH POS TAGGER

4. TỐI ƯU HÓA POS TAGGER

5.NHẬN DIỆN THỰC THỂ CHƯA ĐẶT TÊN (UNER)

6. NHẬN BIẾT TỪ QUAN HỆ

7. XÁC ĐỊNH LỚP CỦA CÂU HỎI

8.GOM THỰC THỂ

9.NHẬN DẠNG QUAN HỆ ẨN

10.XÁC ĐỊNH LOẠI QUAN HỆ GIỮA CÁC THỰC THỂ

11. XÓA BỎ QUAN HỆ KHÔNG PHÙ HỢP

12. XÂY DỰNG ĐỒ THỊ Ý NIỆM

13.RÚT CÂU TRUY VẤN

**1.Nhận diện thực thể đặt tên**

Theo ý kiến đề xuất của nhóm, trong câu, thực thể đặt tên là những từ được nhóm quy định bởi quy tắc sau:

* Những từ viết hoa hoặc những cụm từ viết hoa liên tiếp.

Ví dụ: Mark Twain,John,…là những từ (cụm từ) dùng để xác định thực thể đặt tên.

* Những từ hoặc cụm từ trong dấu ngoặc kép <“>.

Ví dụ: Who wrote “I am Sam”? Như vậy cụm “I am Sam” là một cụm từ dùng để xác định thực thể đặt tên.

Vì những câu hỏi chỉ tập trung vào những gì Ontology nên thực thể đặt tên được nhận dạng bằng cách truy xuất Ontology tương ứng để tìm lớp tương ứng.

Ví dụ: Cụm từ *“Mark Twain”* khi truy xuất vào Ontology thì sẽ tìm thấy lớp tương ứng chứa nó là **Author**. Như vậy phân lớp tương ứng của cụm từ *“Mark Twain”* là **Author**.

**2. Tiền xử lý Pos Tagger**

Xử lí Pos Tagger là quá trình xác định từ loại của một từ trong câu. Hiện nay, có các nhiều công cụ để giúp ta có thể xử lý Pos Tagger được đưa ra như GATE, **OPENNLP** **(cái này có POS Tagger ko Hưng?) ,wordnet** … Mặc dù các công cụ đó đã và đang đáp ứng hầu như mọi dạng câu **(dạng câu ở đây Hưng muốn đề cập là gì ?),** tuy nhiên các công cụ đó **chưa phải chính xác hoàn toàn( có ai đánh giá như vậy ?).**

Điển hình: Các công cụ đó chưa có thể hiểu được những thực thể đặt tên là những cụm từ viết hoa liên tiếp hay những cụm từ trong dấu <”> hoặc <’> đã nêu mà thay vào là hiểu những từ riêng lẽ trong những cụm từ đó. Như vậy nó đã giảm đi một phần độ chính xác trong hệ thống truy xuất Ontology. Vì thế, nhóm đề xuất quy tắc tiền xử lí Pos Tagger như sau:

Thay những cụm từ của những thực thể đặt tên đã xác định trước đó bằng một dãy các ký tự bắt đầu bằng chữ hoa tương ứng với mã số kèm theo. Như vậy, các công cụ xử lí Pos Tagger sẽ hiểu những cụm từ đó là các NNP.

Cụ thể, các thực thể đặt tên đó sẽ thay bằng công thức : NER+MÃ SỐ

Ví dụ: Với câu hỏi Who wrote “I am Sam”, kết quả xử lí Pos Tagger như sau:

* Chưa tiền xử lý:

*WP VBD " PRP  VBP NNP " ?*

* Đã tiền xử lý, đặt “I am Sam” là NER001

WP VBD NNP ?

Trong đó:

WP: Đại từ câu hỏi.

VBD: Động từ dạng quá khứ.

NNP:Danh từ viết hoa số ít.

PRP: Đại từ chỉ người.

**=> bước trên đã nhận diện thực thể đặt tên rồi thì ở bước này có cần thiết gom các từ thuộc NNP lại không ? hoàng thấy bước này đơn giản chỉ cần nhận dạng từ loại thôi .**

**3. Xây dựng danh sách Pos Tagger**

Việc xử lý Pos Tagger sẽ giúp ta dễ dàng xác định từ loại trong câu, từ đó dễ dàng ánh xạ với các bộ từ điển ngữ nghĩa như Wordnet. Hơn nữa nó còn giúp ta xác định các từ quan hệ và các từ mà sau này dùng để xác định thực thể chưa đặt tên.

**4. Tối ưu hóa danh sách Pos Tagger**

Mặc dù đã tiền xử lí ở bước trên, tuy nhiên khi xác định động từ trong câu thì các kết quả ở trên vẫn chưa hề chính xác về mặt ngữ nghĩa của nó.

Ví dụ: động từ “look up” là sự kết hợp giữa động từ “look” và giới từ “up”. Tuy nhiên, nếu chỉ dựa vào các bước xử lý ở trên để xác định động từ thì kết quả cho ra lại là “look”. Điều này không thật sự chính xác.

Do vậy, nhóm đề xuất phương pháp tối ưu hóa danh sách Pos Tagger cùng với sự hỗ trợ của bộ từ điển Wordnet để xác định nghĩa của từ, với các bước sau:

* Chuyển mẫu câu về dạng dạng hiện tại đơn.
* Nếu

+ *Động từ* + *Giới từ* là một cụm từ có nghĩa:

Động từ = Động từ + Giới từ

+ Ngược lại: Động từ = Động từ.

**=> bước này có lý , nhưng với CSDL mới là MySQL thì chắc ko cần xét tới . hoặc liệt kê một danh sách các từ quan hệ .**

**Vd : write, look up ,is author of , related to,….**

**5. Nhận diện thực thể chưa đặt tên**

Việc nhận dạng thực thể chưa đặt tên sẽ giúp ta tìm ra được thực thể mà Ontology đã tồn tại gần hoặc cùng nghĩa nhưng khác từ. Ví dụ

Từ “writer” sẽ đồng nghĩa với từ “author” . Nếu như không nhận dạng thì ta sẽ không trả lời được những câu hỏi liên quan đến writer.

Nhóm đề xuất sử dụng công cụ Similarity Wordnet để nhận dạng thực thể chưa đặt tên. Similarity Wordnet là một công cụ dùng để đánh giá độ tương đồng về nghĩa giữa 2 danh từ. Như vậy, với cách này ta có thể chọn lớp thực thể trong Ontology có độ tương đồng về nghĩa lớn nhất để làm lớp cho thực thể chưa đặt tên.

Ví dụ: Với câu hỏi What is the author of “Harry Porter”?

Thực thể chưa đặt tên ở đây (ngoại trừ từ what) là author. Giả sử trong Ontology đã có chứa một lớp có tên là Writer.

Bằng công cụ Similarity Wordnet, ta có thể chọn lớp cho từ *author* là *writer* thay vì là *Book* hay bất kỳ tên lớp nào khác trong Ontology.

**=> đồng ý**

**6. Nhận biết từ quan hệ**

Các từ quan hệ là yếu tố cần thiết cho việc xác định mỗi quan hệ giữa các thực thể trong câu truy vấn. Các từ quan hệ là các động từ hoặc giới từ trong câu như of,in, at, write….

**=> cái này có giống bước 4 ko ? hay bước 4 chỉ gom lại rồi đến bước 6 mới nhận biết từ quan hệ .**

**7. Xác định lớp của câu hỏi:**

- Đối với câu hỏi What:

+ Với dạng câu hỏi *What do/does +Noun + mean?* Thì lớp câu hỏi chính là lớp của danh từ.

+ Với dạng câu hỏi *What is/are +Noun+ stand/short for?* Thì lớp câu hỏi là lớp trong Ontology có độ tương đồng về nghĩa lớn nhất với từ “name”.

+ Với dạng câu hỏi : *What is/are + Noun?* thì lớp câu hỏi chính là lớp của danh từ .

+ Với dạng câu hỏi *What+ Noun+….?* Thì lớp câu hỏi chinh là lớp của danh từ.

+ Với những trường hợp còn lại, thì việc xác định lớp của What phải dựa vào lớp của danh từ có quan hệ với What và từ quan hệ giữa chúng và tiến hành sao khớp 2 thông tin đã biết đó vào Ontology để tìm lớp của What.

-Đối với câu hỏi Who:

+Với dạng câu hỏi *Who is Noun…?* Thì lớp của câu hỏi chính là lớp của Noun

+Với dạng câu hỏi Who do/does +Noun + Verb thì Lớp của câu hỏi Who chỉ có thể liên quan đến Person hoặc Group, vì thế ta tìm tất cả những lớp có độ tương đồng về nghĩa với “Person” và “Group” đồng thời có mối quan hệ liên quan với động từ Verb.

**-Đối với câu hỏi Which:**

+ Với dạng câu hỏi : Which is/are+Noun…? Thì lớp câu hỏi chính là lớp của Noun

+ Với dạng câu hỏi : Which + Noun …? Thì lớp câu hỏi chinh là lớp của Noun

- Đối với câu hỏi How:

Với dạng câu hỏi : How +adj+…? Bằng cách dùng từ điển Wordnet, ta có thể tìm ra danh từ tương ứng của tính từ (Adj) và lớp của danh từ đó chính là lớp của câu hỏi.

**=> OK !**

**8. Gom thực thể:**

Việc gom thực thể sẽ giúp ta loại bỏ những bộ ba quan hệ không cần thiết, giảm thiểu việc truy vấn dữ liệu.

Theo tài liệu “Dịch câu truy vấn có cấu trúc sang đồ thị ý niệm” của Thạc sĩ Cao Duy Trường thì hai thực thể chỉ được gom lại khi chúng thõa mãn điều kiện sau:

1. Trong 2 thực thể có ít nhật một thực thể chưa đặt tên.
2. Lớp của thực thể không tên là lớp cha của thực thể còn lại xét trên miền Ontology.
3. Giữa 2 thực thể là dấu phẩy <,> hoặc các từ quan hệ đặc biệt : is, are,was,were.

**9. Xác định quan hệ ẩn**

Vì một số trường hợp giữa 2 thực thể trong câu hỏi đứng liền kề nhau nhưng giữa chúng không có từ quan hệ nào, vì thế ta cần xác định quan hệ ẩn giữa chúng.

Ví dụ: Who was born in Ho Chi Minh, VietNam?

Giữa 2 từ Ho Chi Minh và Viet Nam có mối quan hệ ẩn với nhau. Giả sử lớp của Ho Chi Minh là City và lớp của VietNam là Country. Ta cần tìm quan hệ giữa 2 lớp City và Country.

**10. Xác định loại quan hệ giữa các thực thể:**

Ứng với các bộ ba <Thực thể, từ quan hệ, thực thể> , chúng ta cần ánh xạ vào Ontology để sinh ra bộ ba tương ứng là <thực thể, loại quan hệ, thực thể>.

**11.Loại bỏ những quan hệ không phù hợp**

* Đó là những quan hệ dư thừa hoặc đã xác định từ những quan hệ trước

Ví dụ “Xuan Dieu used to grow up in Ha Noi, Viet Nam”.

Sẽ xuất hiện 3 quan hệ:

<Xuân Dieu , grow up, HaNoi>

<Xuân Dieu , grow up, Viet Nam>

<HaHoi,thu do, Viet Nam>

Như vậy khi truy vấn ta chỉ cần truy vấn 2 trong 3, vì 3 thực thể trên đều có quan hệ với nhau và trở thành chu trình.

* Hoặc các quan hệ tạo thành vòng lặp hoặc dạng sau:

<Thực thể > là <Thực thể>. Trong đó 2 thực thể này có cùng loại với nhau.

Ví dụ: He is a author. Trong này 2 thực thể có cùng loại thực thể