**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**--------------** -----🙞🙜🕮🙞🙜----- **------------**



**HỆ THỐNG HỎI ĐÁP TỰ ĐỘNG VỀ LUẬT GIAO THÔNG**

|  |  |
| --- | --- |
| **Giảng Viên Hướng Dẫn:** | **TS. Ngô Xuân Bách** |
| **Sinh Viên Thực Hiện:** | **Lê Thị Ngọc Châm** |
| **Trần Hà Ngọc Thiện** |

Hà Nội, tháng 03, năm 2017

**Mục Lục**

[1 Đặt vấn đề 1](#_Toc477860152)

[1.1 Tổng quan 1](#_Toc477860153)

[1.2 Mô tả bài toán 1](#_Toc477860154)

[2 Giải pháp 1](#_Toc477860155)

[3 Phân tích câu hỏi (Question Analysis) 2](#_Toc477860156)

[3.1 Phương pháp phân tích 2](#_Toc477860157)

[3.1.1 Nhận diện thành phần 2](#_Toc477860158)

[3.1.2 Các nhãn cần nhận diện 3](#_Toc477860159)

[3.2 Huấn luyện sử dụng CRF 3](#_Toc477860160)

[4 Dữ liệu 5](#_Toc477860161)

[5 Thực nghiệm 6](#_Toc477860162)

[6 Đánh giá kết quả 9](#_Toc477860163)

# Đặt vấn đề

## Tổng quan

Hệ thống hỏi đáp tự động là hệ thống được xây dựng để thực hiện việc tìm kiếm tự động câu trả lời từ một tập lớn các tài liệu cho câu hỏi đầu vào một cách chính xác. Từ những năm 1960, các hệ thống hỏi đáp đầu tiên đã được ra đời. Điểm chung trong các hệ thống này là sử dụng cơ sở dữ liệu được thiết kế bằng tay bởi các chuyên gia trong lĩnh vực được chọn [MJ08]. Giai đoạn những năm 1970- 1980, có nhiều dự án lớn hướng đến việc “hiểu văn bản” và xây dựng hệ thống hỏi đáp dựa trên các mô hình ngôn ngữ thống kê. Hội nghị TREC diễn ra hàng năm (bắt đầu từ cuối những năm 1990) thu hút sự tham gia của rất nhiều các nhóm nghiên cứu cũng đã góp phần rất lớn trong việc thúc đẩy các nghiên cứu về hệ thống hỏi đáp. Cuối những năm 1990, world wide web ra đời và nhanh chóng phát triển bùng nổ trở thành một kho ngữ liệu khổng lồ. Các nhà nghiên cứu về hệ thống hỏi đáp cũng bắt đầu khai thác web như là một nguồn cho việc tìm kiếm câu trả lời. Các kĩ thuật mới đòi hỏi tốc độ cao, khả năng xử lý lượng dữ liệu web lớn đang rất được quan tâm.

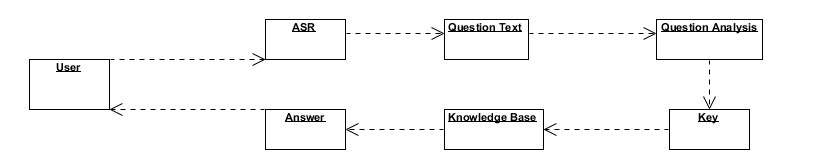
## Mô tả bài toán

Bài toán đặt ra là khi có người hỏi một câu hỏi về luật giao thông với hệ thống, hệ thống sẽ tự động phân tích câu hỏi từ đó trích chọn thông tin để tìm và đưa ra câu trả lời cho câu hỏi.

# Giải pháp

Giải pháp được đưa ra đối với hệ thống bao gồm các bước theo thứ tự như sau:

* Nhận dạng tiếng nói (**Auto Speech Recognition**): xứ lý tiếng nói để đưa được nội dung câu hỏi về dạng văng bản.
* Phân tích câu hỏi (**Question Analysis**): sau khi câu hỏi được đưa về dạng văn bản, ta tiến hành phân tích câu hỏi để có thể tìm ra được các từ khóa, nội dung quan trọng của câu hỏi.
* Tìm kiếm trong cơ sở tri thức (**Knowledge Base**): dựa trên các từ khóa và nội dung quan trọng của câu hỏi, tiến hành tìm kiếm trong cơ sở tri thức và đưa ra câu trả lời phù hợp.



# Phân tích câu hỏi (Question Analysis)

## Phương pháp phân tích

### Nhận diện thành phần

Để trả lời được câu hỏi, ta càn nhận diện được các từ khóa, thành phần quan trọng trong câu. Cụ thể với bài toán trả lời câu hỏi về luật giao thông ta cần phải nhận diện được các thông tin quan trọng như phương tiện là gì, phương tiện ra làm sao, trường hợp, thông tin hương tới trong câu hỏi, …

Các từ khóa và các thành phần quan trọng trong câu sẽ được nhận diện bằng phương pháp học máy. Máy tính sẽ tiến hành luyện tập và tự động nhận diện và gán nhãn các từ khóa, các thành phần quan trọng.

Như vậy, cần phải có một tập dữ liệu huấn luyện được gán nhãn sẵn để máy có thể tiến hành huấn luyện nhận diện các thành phần. Trong đó, với mỗi một câu thì chỉ có các từ khóa và các thành phần quan trọng của câu được gán nhãn nhận diện.

Cấu trúc gán nhãn cho một từ khóa hay một thành phần quan trọng như sau: .

Trong đó là nhãn và content là nội dung được gán nhãn.

Ví dụ gán nhãn câu như sau:

* Trước gán nhãn: *đi ô tô không nhường đường cho xe thô sơ khi chuyển hướng phạt bao nhiêu tiền?*

Sau gán nhãn: *đi <tv>ô tô</tv> <a>không nhường đường cho xe thô sơ khi chuyển hướng</a> <qt>phạt bao nhiêu tiền</qt>?*

* Trước gán nhãn: *điều khiển xe ô tô không tuân thủ quy định khi ra vào đường cao tốc bị phạt bao nhiêu tiền?*

Sau gán nhãn: *điều khiển xe <tv>ô tô</tv> <a>không tuân thủ quy định</a> <if4>khi ra vào đường cao tốc</if4> <qt>bị phạt bao nhiêu tiền</qt>?*

Dựa trên tập dữ liệu như vậy, máy sẽ tiến hành học để có thể nhận diện ra các thành phần quan trọng.

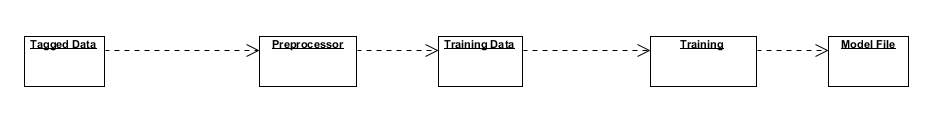
### Các nhãn cần nhận diện

Các từ khóa, thành phần quan trọng của câu được chia thành nhiều loại và có ý nghĩa khác nhau trong từng trường hợp vì vậy cũng cần có các nhãn khác nhau tương ứng. Cụ thể với bài toán trả lời về luật giao thông, ta có các nhãn:

|  |  |
| --- | --- |
| Tên Nhãn | Ý Nghĩa |
| tp | Người tham gia giao thông |
| tv | Loại phương tiện |
| if1 | Thông tin về phương tiện |
| a | Hành động của phương tiện |
| l | Vị trí |
| tr | Loại đường |
| ac | Nồng độ cồn |
| sp | Tốc độ |
| v | Giá trị |
| tl | Tín hiệu giao thông |
| if2 | Thông tin về tín hiệu giao thông |
| ti | Người hướng dẫn giao thông |
| if3 | Thông tin về người tham gia giao thông |
| if4 | Thông tin chung chung |
| qt | Loại câu hỏi |
| ano | Định nghĩa |

## Huấn luyện sử dụng CRF

Sơ đồ tiến hành huấn luyện:



Tập dữ liệu đã được gán nhãn sẽ được tiền xử lý trước khi được đưa vào huấn luyện. Cụ thể các token trên mỗi câu sẽ được tách ra đặt trên một dòng cùng với nhãn từ loại và nhãn được gắn của token đó, các câu sẽ cách nhau bằng một dòng trống. Những token bắt đầu của nhãn sẽ được đánh theo cấu trúc B-tag, và những token tiếp theo của nhãn đó sẽ được đánh theo cấu trúc I-tag với tag là tên nhãn.

Ví dụ:



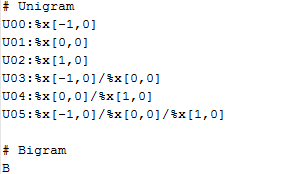
Sau đó tiến hành luyện tập dựa trên tập dữ liệu đã xử lý này và đưa ra file model, là file chứa các thông số sau khi luyện tập để phục vụ cho việc nhận dạng các từ khóa và thành phần quan trọng cho câu hỏi đầu vào.

Khi huấn luyện để tạo ra file model, ta cần phải có file template để quy ước việc nhận dạng một nhãn phụ thuộc vào các thành phần đứng trước và sau như thế nào.

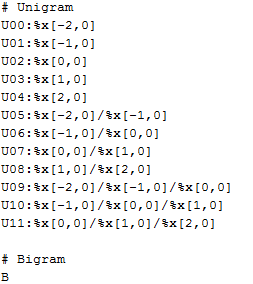
Ví dụ khi muốn phát hiện ở vị trí hiện tại phụ thuộc vào vị trí trước đó ở cột số 0 thì ta có: %x[-1,0].

Ta sẽ huấn luyện học máy trên 3 mô hình đó là 1-gram, 2-gram và 3-gram. Với các file template của 3 mô hình này như sau:

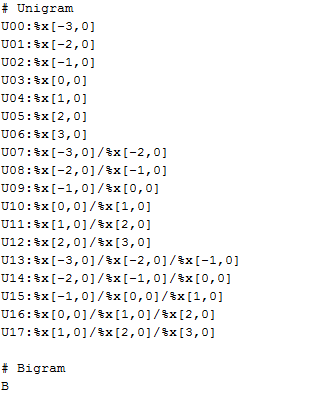
* 1-gram:



* 2-gram:



* 3-gram:



# Dữ liệu

Dữ liệu bao gồm 1628 câu hỏi, được phân loại thành các chủ đề:

* Câu hỏi về vi phạm:
  + Vi phạm về ô tô.
  + Vi phạm về xe máy.
  + Vi phạm của cá phương tiện khác.
* Câu hỏi luật:
  + Hỏi về định nghĩa.
  + Hỏi về tình huống.
  + Hỏi về giấy phép lái xe.
  + Hỏi về quy tắc.

Bộ dữ liệu này sẽ được tiến hành gán nhãn từ loại và tách từng câu đưa về câu trúc như ở trong phần 3.2 đã nêu, bên cạnh đó cũng sẽ có một bộ dữ liệu như vậy nhưng không tách từ mà mỗi âm tiết được để trên một dòng. Dữ liệu sau khi được đưa về cấu trúc trên, sẽ được tách thành 5 fold. Sau đó tiến hành huấn luyện học máy 5 lần đối vỡi mỗi gram, trong đó mỗi lần lấy 4 fold để huấn luyện và 1 fold còn lại để tiến hành test.

# Thực nghiệm

Kết quả thực nghiệm trung bình sau khi đã chạy 5 lần với mỗi mô hình gram:

* 1-gram:

Với dữ liệu có tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 86.74% | 87.66% | 87.20% |
| ac | 92.55% | 88.91% | 90.55% |
| ano | 81.82% | 58.45% | 67.28% |
| dl | 98.64% | 98.71% | 98.64% |
| if1 | 94.07% | 86.42% | 89.96% |
| if2 | 100.00% | 75.00% | 81.33% |
| if3 | 88.50% | 76.06% | 81.69% |
| if4 | 86.29% | 71.37% | 77.95% |
| l | 85.88% | 84.78% | 85.24% |
| qt | 94.70% | 93.52% | 94.11% |
| sp | 94.37% | 95.67% | 94.97% |
| ti | 100.00% | 96.60% | 98.21% |
| tl | 100.00% | 56.86% | 67.54% |
| tp | 100.00% | 85.33% | 91.00% |
| tv | 96.27% | 97.42% | 96.84% |
| v | 98.61% | 96.94% | 97.74% |
| Overall | 92.44% | 90.17% | 91.29% |

Với dự liệu không tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 85.46% | 87.15% | 86.29% |
| ac | 96.85% | 92.99% | 94.80% |
| ano | 84.41% | 56.54% | 66.96% |
| dl | 98.04% | 97.89% | 97.92% |
| if1 | 92.21% | 82.77% | 87.06% |
| if2 | 80.00% | 70.00% | 73.33% |
| if3 | 85.26% | 72.57% | 78.25% |
| if4 | 82.54% | 67.19% | 74.01% |
| l | 85.95% | 84.08% | 84.97% |
| qt | 94.13% | 92.95% | 93.53% |
| sp | 91.48% | 92.88% | 92.00% |
| ti | 100.00% | 96.93% | 98.42% |
| tl | 81.31% | 59.89% | 68.47% |
| tp | 95.00% | 72.00% | 79.33% |
| tv | 96.39% | 97.37% | 96.88% |
| v | 98.50% | 98.32% | 98.40% |
| Overall | 91.74% | 89.75% | 90.73% |

* 2-gram:

Với dữ liệu có tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 87.97% | 88.16% | 88.06% |
| ac | 98.00% | 94.36% | 96.00% |
| ano | 79.77% | 62.56% | 69.15% |
| dl | 98.64% | 97.97% | 98.26% |
| if1 | 93.43% | 84.42% | 88.61% |
| if2 | 100.00% | 85.00% | 88.00% |
| if3 | 87.41% | 78.89% | 82.85% |
| if4 | 84.08% | 70.09% | 76.26% |
| l | 86.54% | 84.54% | 85.45% |
| qt | 94.69% | 93.36% | 94.02% |
| sp | 96.23% | 95.67% | 95.92% |
| ti | 100.00% | 97.77% | 98.85% |
| tl | 100.00% | 54.86% | 66.98% |
| tp | 94.29% | 83.33% | 87.76% |
| tv | 96.86% | 96.37% | 96.60% |
| v | 97.61% | 97.19% | 97.40% |
| Overall | 92.68% | 90.13% | 91.38% |

Với dự liệu không tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 86.31% | 87.62% | 86.94% |
| ac | 92.99% | 92.99% | 92.99% |
| ano | 88.44% | 61.48% | 71.64% |
| dl | 98.69% | 97.89% | 98.26% |
| if1 | 94.64% | 82.73% | 87.96% |
| if2 | 93.33% | 90.00% | 89.33% |
| if3 | 86.34% | 75.63% | 80.47% |
| if4 | 84.12% | 68.53% | 75.47% |
| l | 86.02% | 83.17% | 84.55% |
| qt | 94.81% | 93.53% | 94.16% |
| sp | 94.96% | 91.54% | 92.94% |
| ti | 99.13% | 96.93% | 98.00% |
| tl | 82.28% | 61.67% | 70.00% |
| tp | 84.33% | 68.00% | 72.67% |
| tv | 97.31% | 97.81% | 97.56% |
| v | 98.30% | 97.32% | 97.80% |
| Overall | 92.45% | 90.20% | 91.31% |

* 3-gram:

Với dữ liệu có tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 87.38% | 87.20% | 87.29% |
| ac | 98.00% | 94.36% | 96.00% |
| ano | 78.53% | 63.67% | 69.42% |
| dl | 99.35% | 97.97% | 98.63% |
| if1 | 93.32% | 83.42% | 88.01% |
| if2 | 100.00% | 75.00% | 81.33% |
| if3 | 88.10% | 78.53% | 82.96% |
| if4 | 85.29% | 71.52% | 77.56% |
| l | 85.74% | 84.01% | 84.76% |
| qt | 94.29% | 93.42% | 93.85% |
| sp | 95.52% | 95.67% | 95.54% |
| ti | 100.00% | 93.93% | 96.73% |
| tl | 100.00% | 47.71% | 61.67% |
| tp | 97.14% | 85.33% | 90.12% |
| tv | 97.01% | 96.05% | 96.52% |
| v | 97.51% | 95.77% | 96.61% |
| Overall | 92.48% | 89.70% | 91.07% |

Với dữ liệu không tách từ:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| a | 87.44% | 88.17% | 87.79% |
| ac | 92.94% | 91.99% | 92.45% |
| ano | 88.71% | 61.48% | 71.73% |
| dl | 98.69% | 97.89% | 98.26% |
| if1 | 95.06% | 80.65% | 86.86% |
| if2 | 93.33% | 90.00% | 89.33% |
| if3 | 86.69% | 79.67% | 82.93% |
| if4 | 85.48% | 69.63% | 76.69% |
| l | 86.65% | 84.24% | 85.42% |
| qt | 94.77% | 93.53% | 94.14% |
| sp | 91.63% | 91.54% | 91.32% |
| ti | 100.00% | 97.65% | 98.79% |
| tl | 84.28% | 61.67% | 70.89% |
| tp | 95.00% | 68.00% | 77.11% |
| tv | 97.32% | 97.81% | 97.57% |
| v | 97.97% | 96.99% | 97.47% |
| Overall | 92.75% | 90.52% | 91.62% |

# Đánh giá kết quả

Dựa vào kết quả trên ta có thể thấy được rằng mô hình 3-gram với dữ liệu ko tách từ đạt kết quả trung bình cao nhất (91.62%). Bên cạnh đó có các nhãn cho kết quả thấp như “tl”, “tp”, “ano”, “if4”.