**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](#_Toc483293527)

[1.1 Tổng quan 1](#_Toc483293528)

[1.2 Mô tả bài toán 1](#_Toc483293529)

[2 Giải pháp 1](#_Toc483293530)

[3 Phân tích câu hỏi (Question Analysis) 2](#_Toc483293531)

[3.1 Phương pháp phân tích 2](#_Toc483293532)

[3.1.1 Nhận diện thành phần 2](#_Toc483293533)

[3.1.2 Các nhãn cần nhận diện 4](#_Toc483293534)

[3.1.3 Phương pháp phân tích câu hỏi 5](#_Toc483293535)

[3.2 Giải pháp 5](#_Toc483293536)

[3.2.1 Giới thiệu về CRF 5](#_Toc483293537)

[3.2.2 Giai đoạn huấn luyện 6](#_Toc483293538)

[3.2.3 Sử dụng CRF 6](#_Toc483293539)

[4 Dữ liệu 8](#_Toc483293540)

[4.1 Nguồn dữ liệu 8](#_Toc483293541)

[4.2 Xây dựng bộ câu hỏi 8](#_Toc483293542)

[4.3 Phân loại câu hỏi 8](#_Toc483293543)

[4.4 Thống kê số lượng mẫu cho mỗi nhãn 11](#_Toc483293544)

[4.5 Đánh giá độ đồng nhất gán nhãn 11](#_Toc483293545)

[4.5.1 Độ đo Cohen’s Kappa 11](#_Toc483293546)

[4.5.2 Độ tương đồng phân cụm 11](#_Toc483293547)

[5 Thực nghiệm 12](#_Toc483293548)

[5.1 Phương pháp CRF 12](#_Toc483293549)

[5.2 Phương pháp HMM 13](#_Toc483293550)

[5.3 So sánh kết quả của 2 phương pháp 14](#_Toc483293551)

[6 Đánh giá kết quả 14](#_Toc483293552)

[6.1 Nhận xét kết quả 14](#_Toc483293553)

[6.2 Phân tích lỗi 14](#_Toc483293554)

# 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.

Hiện nay, hệ thống trả lời câu hỏi cho người dùng còn ít, và cũng chỉ tập trung về một số vấn đề xã hội. Vì vậy, đề tài hệ thống hỏi đáp về luật giao thông này nhằm mục đích xây dựng ra một hệ thống để có thể giúp cho người dùng dùng dù ở bất cứ đâu đều có thể dễ dàng tìm được các thông tin về giao thông, luật giao thông, hay bất cứ thắc mắc gì liên quan tới giao thông bằng việc gửi câu hỏi lên cho hệ thống.

## Mô tả bài toán

Tình huống được đưa ra là một người đang đi trên đường và bỗng có thắc mắc gì đó liên quan tới luật giao thông. Như vậy người đó sẽ nói câu hỏi của mình vào điện thoại, câu hỏi đó sẽ được điện thoại nhận dạng từ tiếng nói và gửi lên cho hệ thống và hệ thống sẽ gửi câu trả lời trở về cho điện thoại để hiện thị.

# Giải pháp

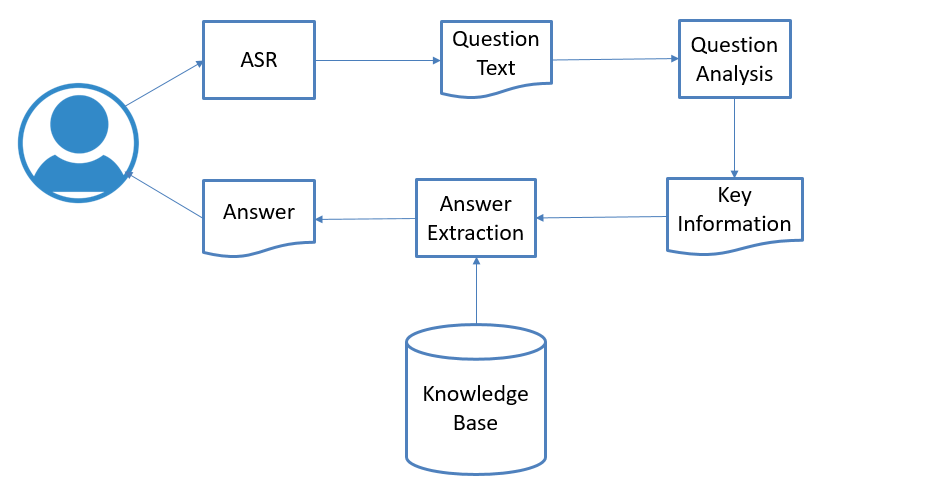
Mục tiêu của hệ thống chính là đưa ra được câu trả lời thỏa mãn được yêu cầu từ câu hỏi của người dùng.

**Ví dụ:**

* ***Câu hỏi***: người lái xe máy thì bị giới hạn nồng độ cồn là bao nhiêu?
* ***Trả lời***: không được vượt quá 50 miligam/100 mililít máu
* ***Câu hỏi***: chuyển hướng ô tô mà không nhường đường cho người đi bộ thì bị phạt bao nhiêu tiền?
* ***Trả lời***: 100.000 đồng đến 200.000 đồng
* ***Câu hỏi***: khái niệm đường bộ được hiểu như thế nào là đúng?
* ***Trả lời***: đường bộ gồm đường, cầu đường bộ. hầm đường bộ, bến phà đường bộ

Quá trình xử lý của hệ thống:

* 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 (**Answer Extraction**): 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 cần nhận diện được các từ khóa và các 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 như thế nào, trong hoàn cảnh nào, loại câu hỏi gì…

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, tự động nhận diện và gán nhãn các từ khóa và các thành phần quan trọng trong câu.

Như vậy, cần phải có một tập dữ liệu huấn luyện đã 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 tên của nhãn và content là nội dung được gán nhãn.

Ví dụ gán nhãn câu như sau:

* Trước khi 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>?*

How much are motorists fined for failing to give away to rudimentary vehicles when changing lanes?

* 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>?*

How much are motorists fined for failing to follow the regulations when entering the highway?

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 | Ví dụ |
| A | Hành động của phương tiện | Vượt, dừng, đỗ |
| AC | Nồng độ cồn | Nồng độ cồn |
| ANO | Định nghĩa | Quy định |
| DL | Bằng lái xe | Giấy phép hạng a1, a2 |
| IF1 | Thông tin về phương tiện | Quái tải trọng, 30 chỗ ngồi |
| IF2 | Thông tin về tín hiệu giao thông | Màu đỏ, tín hiệu vàng |
| IF3 | Thông tin về người tham gia giao thông | Đội mũ bảo hiểm, lái xe |
| IF4 | Thông tin chung chung | Giờ, đường dốc, giơ tay |
| L | Vị trí | Ngã tư, ngã ba, chân cầu vượt |
| QT | Loại câu hỏi | Phạt thế nào, phạt bao nhiêu |
| SP | Tốc độ | Tốc độ |
| TI | Người hướng dẫn giao thông | Cảnh sát |
| TL | Tín hiệu giao thông | Đèn xanh, đèn đỏ, biển báo |
| TP | Người tham gia giao thông | Người đi bộ, hành khách |
| TV | Loại phương tiện | Ô tô, xe máy, xe đạp |
| V | Giá trị | 10km/h |

|  |  |  |
| --- | --- | --- |
| Tags’ name | Mean | Example |
| A | Action of vehicle | Overtake, reverse |
| AC | Acohol concentration | Acohol concentration |
| ANO | Annotation | Motor way, main road |
| DL | Driving licence | Licence class A1 |
| IF1 | Information about means of traffic | 30 seats |
| IF2 | Information about traffic lights | Red, yellow |
| IF3 | Information about traffic participants | Wear helmet |
| IF4 | More information | Bypass, ring road |
| L | location | Crossroads, fork |
| QT | Type of questions | How much |
| SP | Speed | speed |
| TI | Traffic instructors | Traffic police |
| TL | Traffic lights | signpost |
| TP | Traffic participants | Pedestrian, motorists |
| TV | Type of vehicle | Car, motorbike |
| V | Value | 10km/h |

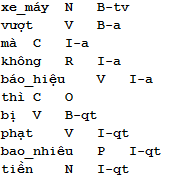
### Phương pháp phân tích câu hỏi

Bài toán được chuyển thành việc gán nhãn cho chuỗi. Mỗi thành phần (từ, âm tiết) được gán một trong các nhãn như B-TV, I-TV, B-TP, I-TP, …

Trong đó:

* B- đánh dấu bắt đầu cụm được gán nhãn
* I- đánh dấu thành phần tiếp theo của cụm được gán nhãn

Ví dụ:





Thành phần là âm tiết Thành phần là từ

## Giải pháp

### Giới thiệu về CRF

***CRF*** là một thực hiện mã nguồn mở đơn giản có thể tùy biến của các trường ngẫu nhiên có điều kiện (***Conditional Random Fields\_ CRFs***) để phân đoạn hoặc gán nhãn các dữ liệu tuần tự.

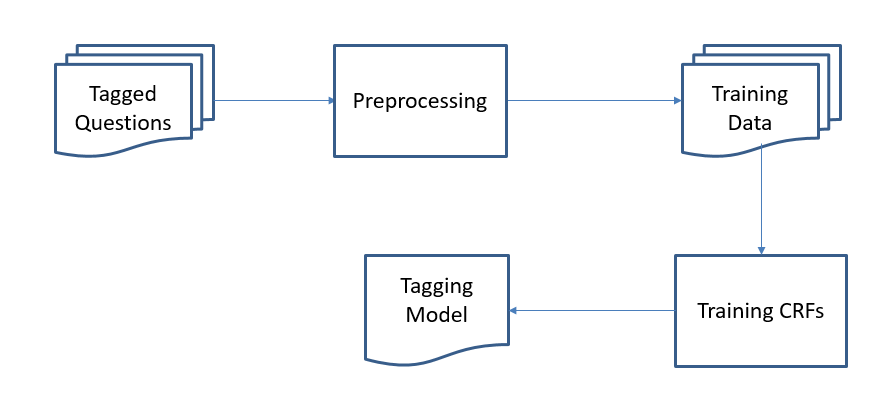
***CRF*** được ứng dụng nhiều trong lĩnh vực NLP như nhận dạng thực thể tên (Named Entity Recognition), trích xuất thông tin (Information Extraction),…

Một số đặc trưng của ***CRF***:

* Được viết bằng C++ với thư viện STL
* Sử dụng ít bộ nhớ cả trong training và test
* Mã hóa và giải mã trong thời gian thực

### Giai đoạn huấn luyện

Sơ đồ huấn luyện

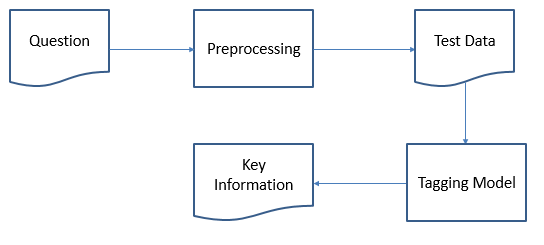


Tập dữ liệu đã được gán nhãn sẽ được tiền xử lý trước khi được đưa vào huấn luyện. Cụ thể các thành phần(từ hoặc âm tiết) trên mỗi câu sẽ được tách ra đặt trên một dòng cùng với nhãn từ loại và nhãn được gắn của thành phần đó, các câu sẽ cách nhau bằng một dòng trống. Những thành phần bắt đầu của nhãn sẽ được đánh theo cấu trúc B-tag, và những thành phần tiếp theo của nhãn đó sẽ được đánh theo cấu trúc I-tag với tag là tên nhãn(cấu trúc như mục 3.1.3).

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.

### Sử dụng CRF

Sơ đồ gán nhãn



***Các đặc trưng sử dụng***

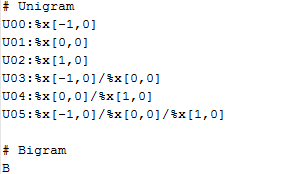
* Sử dụng mức từ và âm tiết
* Từ là đơn vị nhỏ nhất, cấu tạo ổn định, mang nghĩa hoàn chỉnh, được dùng để cấu thành nên câu: Ô tô, xe máy, đi, dừng, giấy phép,…
* Âm tiết là thành phần cấu tạo nên từ: Ô, tô, phạt, bao, nhiêu,…
* Kích thước cửa sổ (window size)

Trước khi huấn luyện để tạo ra file model, ta cần phải quy ước việc nhận dạng một nhãn phụ thuộc vào các thành phần đứng trước và sau như thế nào. Những quy ước này được chứa trong các file template.

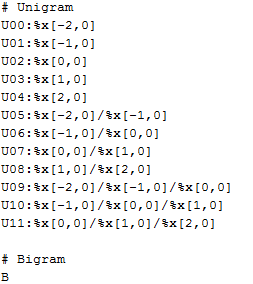
***Ví dụ:*** từ đang xét ở vị trí x, sử dụng 2 từ ở vị trí x-1 và x+1 để tham gia gán nhãn từ x. Tổng cộng là 3 từ ta có kích thước cửa sổ là 3

Ta sẽ huấn luyện học máy trên 3 mô hình có window size lần lượt là 3, 5, 7 và chỉ sử dụng Unigram. Định dạng các file template của 3 mô hình này như sau:

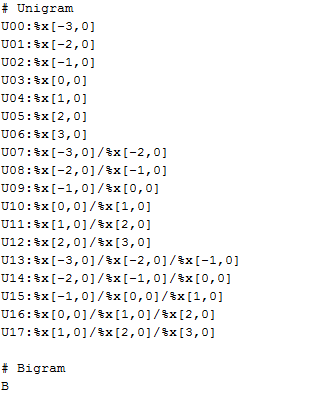
***Window size*** = 3:



***Window size*** = 5:



***Window size*** = 7:



# Dữ liệu

## Nguồn dữ liệu

Các bộ luật về giao thông được lấy từ trang web: [http://thuvienphapluat.vn](http://thuvienphapluat.vn/), [http://vanban.chinhphu.vn](http://vanban.chinhphu.vn/)

## Xây dựng bộ câu hỏi

Hai người đọc nội dung của các nghị định và thông tư về luật giao thông và đặt các câu hỏi dựa vào các nội dung đó.

Bộ câu hỏi về thi bằng lái xe ô tô và xe máy được lấy về từ trang web: <http://vnexpress.net/interactive/2016/thi-sat-hach-lai-xe> và <http://www.thuexemientrung.net/150-cau-hoi-thi-bang-lai-xe-may.html>

Hai người gán nhãn bằng tay các câu hỏi đặt ra sau đó người thứ 3 sẽ kiểm tra lại.

## Phân loại câu hỏi

Các câu hỏi được chia thành 2 loại câu hỏi chính: Vi phạm và hướng dẫn. Trong mỗi loại lại được chia thành các loại nhỏ hơn. Cụ thể:

* Vi phạm: những câu hỏi về việc phạt các lỗi vi phạm.
* Vi phạm của ô tô  
  ***VD***: ô tô không nhường xe thô sơ phạt bao nhiêu tiền?

How much are motorists fined for failing to give way to rudimentary vehicles?

* Vi phạm của xe máy  
  ***VD***: dùng ô khi lái xe máy bị phạt bao nhiêu tiền?

How much are motorcycle riders fined for using the umbrella?

* Vi phạm của phương tiện khác  
  ***VD***: xe đạp không chấp hành biển báo bị phạt bao nhiêu tiền?

How much are bicycle riders fined for failing to follow the signposts?

* Hướng dẫn: những câu hỏi về các hành vi, các trường hợp và các định nghĩa trong khi tham gia giao thông bao gồm.
* Định nghĩa  
  ***VD***: đường bộ trong khu vực đông dân cư gồm những đường nào?

What sort of roads are there in the populated areas?

* Tình huống  
  ***VD***: ban đêm xe cơ giới đi ngược chiều gặp nhau, đèn chiếu sáng phải được sử dụng như thế nào?

How do the drivers use the vehicle lights when facing oncoming traffic at night?

* Giấy phép lái xe  
  ***VD***: điều khiển xe cơ giới có cần phải có giấy phép lái xe không?

Are rudimentary drivers in need of driving licence?

* Quy tắc  
  ***VD***: hành vi điều khiển xe cơ giới giành đường có bị nghiêm cấm hay không?

Do the acts of roads usurpation be strictly forbidden?

Dữ liệu bao gồm 1678 câu hỏi và các câu trả lời kèm theo được chia thành các nhóm câu hỏi như sau:

|  |  |  |
| --- | --- | --- |
| **Loại câu hỏi** | | **Số lượng** |
| Vi phạm | Ô tô | 375 |
| Xe máy | 345 |
| Các phương tiện khác | 152 |
| Hướng dẫn | Định nghĩa | 73 |
| Tình huống | 270 |
| Giấy phép lái xe | 134 |
| Quy tắc | 329 |
| Tổng | | 1678 |

|  |  |  |
| --- | --- | --- |
| **Type of question** | | **Quantity** |
| Fine | Car | 375 |
| Motobike | 345 |
| Others | 152 |
| Instruction | Concept | 73 |
| Situation | 270 |
| Driving Licence | 134 |
| Regulation | 329 |
| Total | | 1678 |

Bộ dữ liệu này sẽ được tiến hành gán nhãn từ loại và sau đó được tiền xử lý trước khi đưa vào huấn luyện. Dữ liệu sau khi tiền xử lý được tách thành 5 fold.

Với mỗi một mô hình huấn luyện, tiến hành huấn luyện học máy 5 lần, mỗi lần lấy 4 fold để huấn luyện và 1 fold còn lại để test. Kết quả thực nghiệm được lưu lại để tiến hành đánh giá các mô hình khác nhau.

## Thống kê số lượng mẫu cho mỗi nhãn

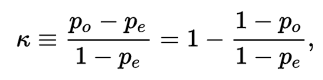
|  |  |
| --- | --- |
| Tên Nhãn | Số lượng mẫu |
| QT | 1678 |
| TV | 1246 |
| A | 1087 |
| L | 426 |
| IF3 | 287 |
| V | 231 |
| IF4 | 227 |
| IF1 | 196 |
| DL | 119 |
| SP | 115 |
| TI | 93 |
| ANO | 75 |
| AC | 44 |
| TL | 31 |
| TP | 20 |
| IF2 | 12 |

## Đánh giá độ đồng nhất gán nhãn

*a. Cơ sở lý thuyết*

Độ đo Cohen’s Kappa được dùng để đo sự đồng thuận giữa 2 người gán nhãn

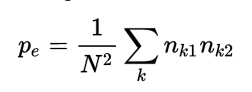
Được tính như sau:



Trong đó:

po: được định nghĩa giống với độ chính xác

pe được tính bằng công thức:



N: số lượng cụm nhãn được gán

nki: số lần người thứ i dự đoán vào nhãn k.

*b. Thực hiện*

Chọn 100 câu hỏi trong bộ dữ liệu sau đó để 2 người tiến hành gán nhãn độc lập. Thực hiện tính độ đo Kappa trên bộ dữ liệu mà 2 người đã gán được và cho kết quả k = 0.84.

# Thực nghiệm

Thực nghiệm sẽ được tiến hành trên bộ dữ liệu đã được xây dựng và sử dụng 2 phương pháp đó là CRF và HMM (Hidden Markov Model) và rút ra kết luận kết quả giữa 2 phương pháp.

## Phương pháp CRF

Kết quả thực nghiệm trung bình sau khi đã chạy 5 lần với các kích thước cửa sổ:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | window size | Precision (%) | Recall | F |
| (%) | (%) |
| Tách từ | 3 | 93.44% | 90.84% | 92.12% |
| 3(pos) | 93.41% | 91.32% | 92.35% |
| 5 | 93.92% | 91.12% | 92.50% |
| 5(pos) | 93.74% | 91.61% | 92.66% |
| 7 | 93.80% | 90.95% | 92.35% |
| 7(pos) | 93.76% | 91.48% | 92.60% |
| Âm tiết | 3 | 93.21% | 91.06% | 92.12% |
| 3(pos) | 93.44% | 91.71% | 92.57% |
| 5 | 93.38% | 91.22% | 92.28% |
| 5(pos) | 93.84% | 92.05% | 92.94% |
| 7 | 93.66% | 91.66% | 92.65% |
| 7(pos) | 93.66% | 92.00% | 92.82% |

Trong đó kết quả trung bình của lần thực nghiệm sử dụng cửa sổ kích thước 5 được tách từ cùng với nhãn từ loại đạt cao nhất (92.94%). Đây là bảng kết quả theo từng nhãn của lần chạy thực nghiệm đó

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F |
| A | 88.20% | 89.14% | 88.66% |
| AC | 95.78% | 95.78% | 95.78% |
| ANO | 85.58% | 60.57% | 68.82% |
| DL | 97.97% | 99.20% | 98.54% |
| IF1 | 94.64% | 87.35% | 90.67% |
| IF2 | 100.00% | 73.33% | 82.67% |
| IF3 | 88.08% | 75.06% | 80.91% |
| IF4 | 85.77% | 74.34% | 79.51% |
| L | 92.23% | 92.71% | 92.44% |
| QT | 96.03% | 94.81% | 95.42% |
| SP | 95.61% | 91.44% | 93.23% |
| TI | 99.05% | 95.53% | 97.24% |
| TL | 86.95% | 60.35% | 68.59% |
| TP | 80.00% | 60.67% | 67.21% |
| TV | 97.32% | 98.78% | 98.04% |
| V | 98.18% | 99.26% | 98.71% |
| Overall | 93.84% | 92.05% | 92.94% |

## Phương pháp HMM

Cùng trên bộ dữ liệu trên, tiến hành huấn luyện và nhận diện trên các bộ test nhưng sử dụng phương pháp HMM.

Tham số được sử dụng khi tiến hành huấn luyện HMM:

* Số lượng chunk được sử dụng để tính lại điểm (Number of underlying chunkings rescored): 64
* Ngram: 12
* Số lượng ký tự trong tập văn bản được sử dụng: 256
* Tỷ lệ nội suy mô hình ngôn ngữ (underlying language-model interpolation ratios): 12
* Kết quả trung bình thu được sau 5 lần chạy bằng phương pháp HMM:

|  |  |  |  |
| --- | --- | --- | --- |
| A | 78.54% | 84.54% | 81.41% |
| AC | 98.00% | 95.78% | 96.82% |
| ANO | 65.97% | 35.56% | 43.69% |
| DL | 99.39% | 98.58% | 98.97% |
| IF1 | 87.83% | 81.70% | 84.56% |
| IF2 | 93.33% | 83.33% | 85.33% |
| IF3 | 83.56% | 78.22% | 80.62% |
| IF4 | 82.29% | 77.57% | 79.77% |
| L | 88.26% | 90.78% | 89.49% |
| QT | 93.95% | 93.63% | 93.79% |
| SP | 87.32% | 90.76% | 88.87% |
| TI | 100.00% | 96.49% | 98.19% |
| TL | 96.67% | 43.27% | 55.96% |
| TP | 60.00% | 20.67% | 30.00% |
| TV | 95.87% | 98.06% | 96.94% |
| V | 97.68% | 99.38% | 98.51% |
| Overall | 89.63% | 89.94% | 89.79% |

## So sánh kết quả của 2 phương pháp

Dựa vào bảng kết quả của 2 phương pháp, ta có thể thấy được CRF cho kết quả tốt hơn HMM rất nhiều, kết quả của CRF trên từng nhãn đều cao hơn so với HMM. Như vậy việc áp dụng phương pháp CRF cho hệ thống sẽ tốt hơn với việc sử dụng phương pháp HMM

# Đánh giá kết quả

## Nhận xét kết quả

Dựa vào kết quả trên ta có thể thấy được rằng mô hình window size = 5 và dữ liệu tách từ đạt kết quả trung bình cao nhất (92.56%). Trong đó, các nhãn “ac, dl, qt, sp, ti, tv, v” cho kết quả cao nhất F>90.00%. Các nhãn “ano, if2, if4, tl, tp” có kết quả thấp hơn cả và F< 80%. CNTTKHMT

## Phân tích lỗi

- *if2, tp* có kết quả thấp hơn do số lượng các nhãn này thấp: *if2* có 12 nhãn, *tp* 20 nhãn

- nhãn *ano* có kết quả thấp do nội dung của nhãn khá giống với các nhãn khác và nội dung thì không thống nhất.

VD:

Câu gốc: <qt>thế nào là</qt> <ano>động cơ 4 kỳ</ano>

CRF: <qt>thế nào là</qt> động cơ <if4>4 kỳ</if4>

Câu gốc: <qt>cơ quan nào quy định</qt> <ano>các đoạn đường cấm đi</ano>

CRF: <qt>cơ quan nào quy định</qt> <if4>các đoạn đường cấm đi</if4>

Câu gốc: <qt>hãy nêu</qt> <ano>công dụng của động cơ ô tô</ano>

CRF: <qt>hãy nêu công dụng của động cơ ô tô</qt>

* Các nhãn bị ngắt quãng trong câu thường bị nhận dạng sai

VD:

Câu gốc: <a>nâng hạng</a> <dl>bằng lái\_xe b 1 số\_tự động</dl> <a>lên b 1</a>

CRF: <a>nâng hạng bằng lái\_xe b 1 số tự\_động lên b</a> <v>1</v>

Câu gốc: <qt>có được</qt> <a>vượt</a> <tl>đèn đỏ</tl> <qt>không</qt>

CRF: <qt>có được vượt đèn đỏ không</qt>

Câu gốc: chỉ có đèn tín hiệu khi <tl>đèn tín hiệu</tl> <if2>màu đỏ</if2> đã bật sáng

CRF: <if4>chỉ có đèn tín hiệu khi đèn tín hiệu màu đỏ đã bật sáng</if4>

Câu gốc: xe <if4>chưa đi qua vạch dừng</if4> trong khi <tl>đèn báo hiệu</tl> <if2>chuyển sang màu vàng</if2> thì xe <qt>có được</qt> <a>đi tiếp</a> <qt>không<qt>

CRF: xe chưa <a>đi qua vạch dừng trong khi đèn báo hiệu chuyển sang màu vàng</a> thì xe <qt>có được đi tiếp không<qt>

* Các câu có nội dung càng dài thì càng khó nhận diện

VD:

Câu gốc: xe <if4>chưa đi qua vạch dừng</if4> trong khi <tl>đèn báo hiệu</tl> <if2>chuyển sang màu vàng</if2> thì xe <qt>có được</qt> <a>đi tiếp</a> <qt>không<qt>

CRF: xe chưa <a>đi qua vạch dừng trong khi đèn báo hiệu chuyển sang màu vàng</a> thì xe <qt>có được đi tiếp không<qt>

Câu gốc: <l>tại nơi đường bộ giao nhau cùng mức với đường sắt</l> chỉ có đèn tín hiệu khi <tl>đèn tín hiệu</tl> <if2>màu đỏ</if2> đã bật sáng, người tham gia giao thông phải dừng lại ngay và <qt>giữ khoảng cách tối thiểu bao nhiêu mét tính từ ray gần nhất</qt>

CRF: <l>tại nơi đường bộ giao nhau cùng mức với đường sắt</l> <if4>chỉ có đèn tín hiệu khi đèn tín hiệu màu đỏ đã bật sáng</if4>, người tham gia giao thông phải <a>dừng lại ngay và giữ khoảng cách tối thiểu bao nhiêu mét tính từ ray gần</a> <qt>nhất</qt>

* Một số trường hợp nhận dạng sai không có lý do:

VD:

Câu gốc: <qt>xe nào</qt> <if1>được quyền ưu tiên đi trước</if1> xe khác

CRF: <qt>xe nào</qt> <a>được quyền ưu tiên đi trước xe khác</a>

Câu gốc: <if3>16 tuổi</if3> thì được lái <qt>xe gì</qt>

CRF: <if3>16 tuổi</if3> thì được lái <tv>xe gì</tv>

* Không nhận dạng được thành phần quan trọng

VD:

Câu gốc: <tv>ô tô</tv> <if1>được kéo</if1> có <qt>cần người lái không</qt>

CRF: <tv>ô tô</tv> được kéo có cần người lái <qt>không</qt>

Câu gốc: <qt>có được phép đi tiếp không</qt>

CRF: <qt>có được</qt> phép đi tiếp <qt>không</qt>