Nhóm 7: Hệ thống dịch câu tiếng việt sang câu tiếng anh

trình bởi Nhóm A

Nghiên cứu khảo sát tổng quan

Trong nền kinh tế hội nhập, thì dịch thuật càng đóng vai trò quan trọng. Hầu hết các lĩnh vực như xây dựng, y tế, tài chính, du lịch, ... đều sử dụng những tài liệu chuyên ngành cả tiếng Việt cũng như ngoại ngữ. Với xu thế hội nhập kinh tế quốc tế thì dịch vụ dịch thuật đối với việc kinh doanh và trao đổi văn hóa là vô cùng thiết yếu.

Một số hệ thống dịch được sử dụng phổ biến hiện nay

- Google Translate
- Linguee
- Bing Translator

Mục tiêu

Ứng dụng công nghệ tiên tiến để thực hiện xây dựng hệ thống dịch tự động các từ,câu tiếng Việt sang tiếng Anh. Hệ thống giúp việc học tập cũng như công việc được dễ dàng hơn. Việc xây dựng hệ thống dịch tự động từ tiếng Việt sang tiếng Anh là rất cần thiết để đáp ứng nhu cầu to lớn của sự phát triển kinh tế và xã hội mang tính chất toàn cầu.

Chương 2 : Xây dựng và thiết kế hệ thống dịch câu tiếng việt sang tiếng anh

2.1.Tổng quan hệ thống

2.1.1.Hệ thống dịch máy thống kê (Statistical Machine Translation)

Với yêu cầu dịch thuật từ tiếng việt sang tiếng anh thì cần đòi hỏi 2 kiến thức chính trong quá trình dịch thuật của dịch thuật viên :

- Lựa chọn từ cho phù hợp với nghĩa của từ ngôn ngữ gốc
- Đúng ngữ pháp (Cấu trúc câu)

Câu hỏi đặt ra là tại sao cần “Lựa chọn từ cho phù hợp với nghĩa” và “Đúng ngữ pháp”. Như chúng ta đã biết, tiếng việt rất đa dạng về thể loại từ cho nên khi ta dịch một từ tiếng việt sang tiếng anh, chúng ta cần phải xem xét sử dụng thể loại từ sao cho đúng với nghĩa gốc.

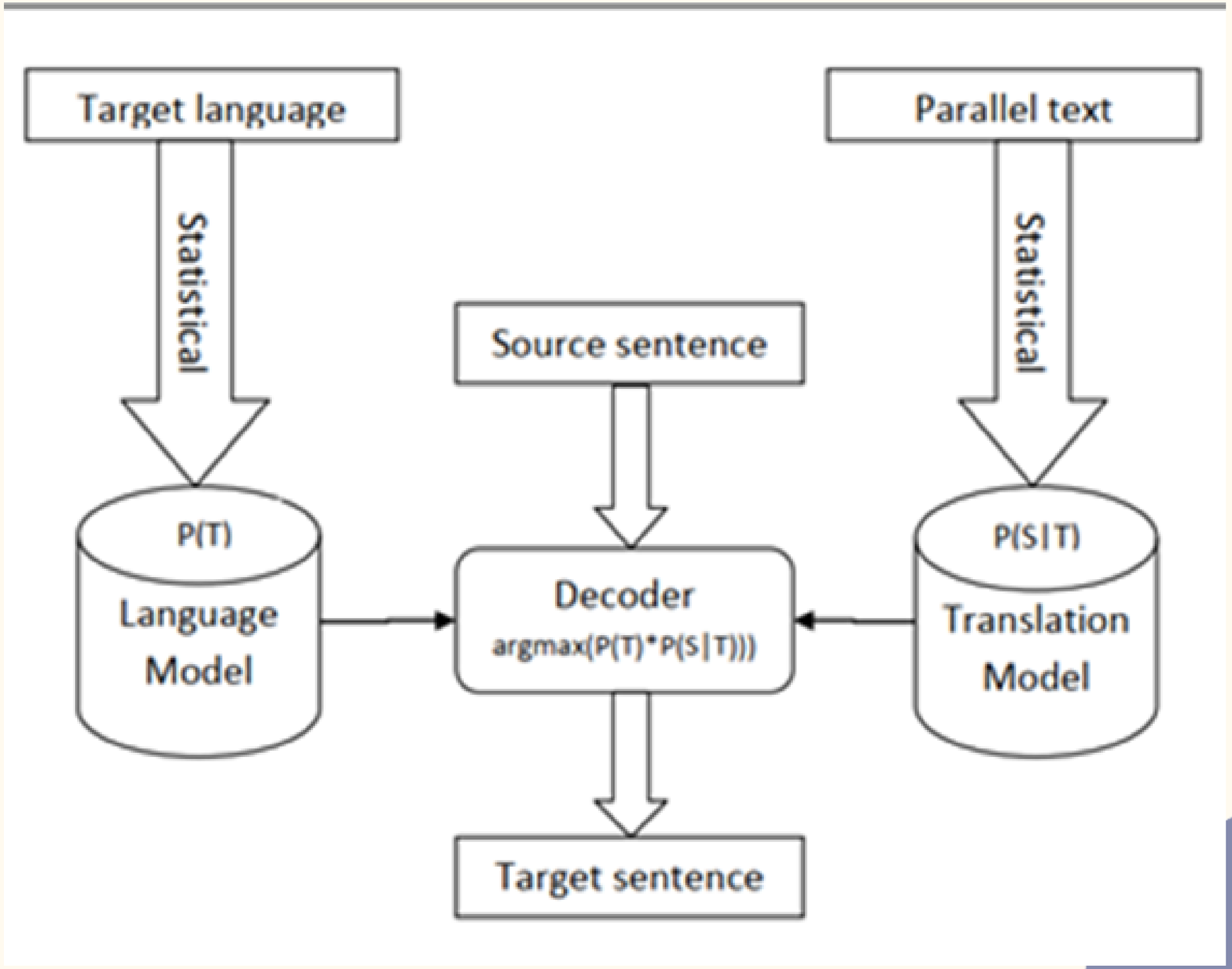
2.1.Tổng quan hệ thống

2.1.1.Hệ thống dịch máy thống kê (Statistical Machine Translation)

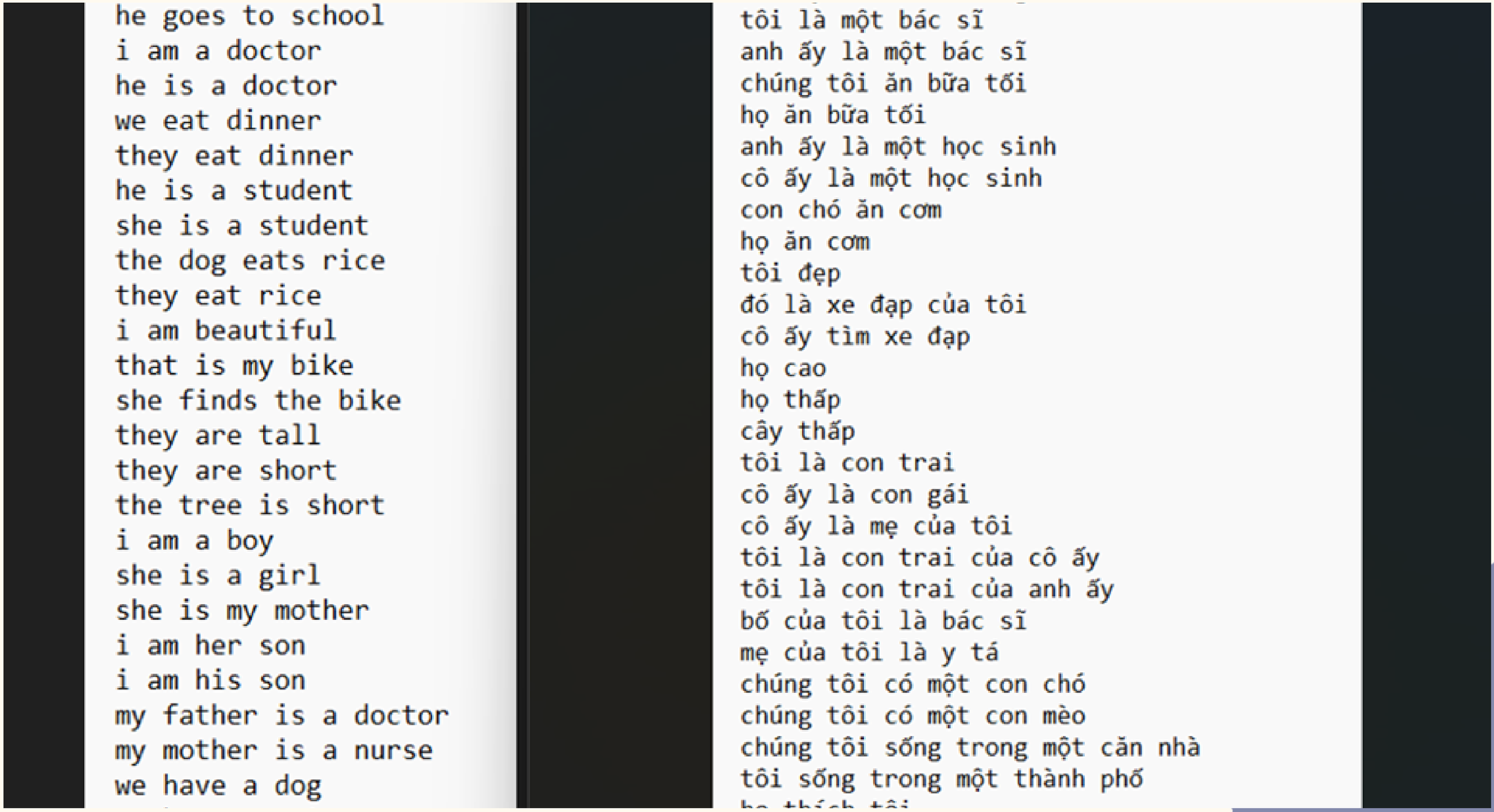
- Từ hai kiến thức chính trong quá trình dịch thuật của dịch thuật viên, người ta đã nghiên cứu và xây dựng ra nhiều hệ thống dịch thuật sử dụng các công nghệ khác nhau. Sau một thời gian tìm hiểu và tham khảo, nhóm em đã quyết định sử dụng Statistical Machine Translation (SMT) để giải quyết bài toán phiên dịch câu tiếng việt sang tiếng anh.

- Mục tiêu cụ thể: Xây dựng hệ thống dịch câu tiếng việt sang câu tiếng anh với chức năng chính:
- Xây dựng hệt thống đưa ra một câu tiếng anh khi người dùng nhập vào một câu tiếng việt
- Kiến trúc tổng quan của hệ thống dịch câu tiếng anh sang tiếng việt được mô tả trong hình 1.

Hình 1- Hệ thống dịch câu tiếng việt sang tiếng anh



2.1.2.Khảo sát các bộ dữ liệu có trong hệ thống dịch câu tiếng việt sang tiếng anh



```
he goes to school
i am a doctor
he is a doctor
we eat dinner
they eat dinner
he is a student
she is a student
the dog eats rice
they eat rice
i am beautiful
that is my bike
she finds the bike
they are tall
they are short
the tree is short
i am a boy
she is a girl
she is my mother
i am her son
i am his son
my father is a doctor
my mother is a nurse
we have a dog
```

```
tôi là một bác sĩ
anh ấy là một bác sĩ
chúng tôi ăn bữa tối
họ ăn bữa tối
anh ấy là một học sinh
cô ấy là một học sinh
con chó ăn cơm
họ ăn cơm
tôi đẹp
đó là xe đạp của tôi
cô ấy tìm xe đạp
họ cao
họ thấp
cây thấp
tôi là con trai
cô ấy là con gái
cô ấy là mẹ của tôi
tôi là con trai của cô ấy
tôi là con trai của anh ấy
bố của tôi là bác sĩ
mẹ của tôi là y tá
chúng tôi có một con chó
chúng tôi có một con mèo
chúng tôi sống trong một căn nhà
tôi sống trong một thành phố
```

2.2.Giải thuật di truyền

Decoder là thành phần chính trong hệ thống SMT. Bản chất của Decoder là một thuật toán tìm kiếm. Vai trò của nó là tìm kiếm bản dịch tốt nhất trong tất cả các bản dịch có thể có. Tuy nhiên, việc sinh ra tất cả các bản dịch và tìm kiếm trong đó là bất khả thi, vì vậy bài toán tìm kiếm chỉ có thể đưa ra lời giải cực trị địa phương. Ở đây, ta sẽ sử dụng giải thuật di truyền làm thuật toán cho Decoder (GADecoder).

2.2.1.Các nghiên cứu liên quan

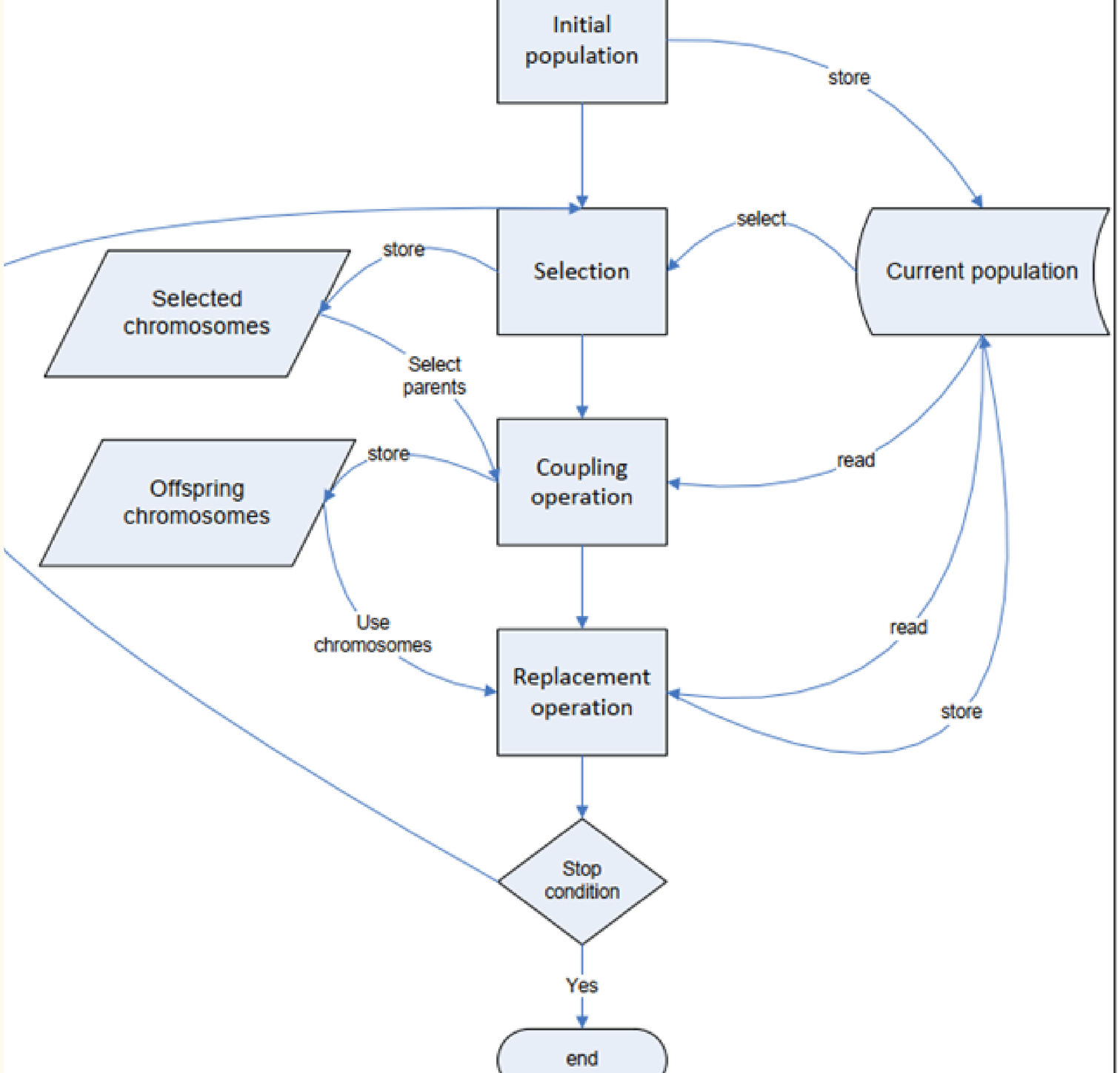
- Về bản chất của hệ thống dịch câu là đi tìm kiếm bản ghi tốt nhất trong một tập hợp gồm rất nhiều bản ghi, đã có rất nhiều phương pháp để giải quyết việc này như Giải thuật Beam Search Decoder, Greedy Decoder, ...
- Giải thuật di truyền (GA) hoạt động với cơ chế tương tự như cơ chế hành vi của gen người. Để khởi tạo quần thể ban đầu, GA tạo ra các NST (nhiễm sắc thể) - lời giải - một cách ngẫu nhiên, trong đó mỗi NST là 1 chuỗi các gen -thành phần tạo nên lời giải. Sau đó, các NST trong quần thể sẽ trải qua quá trình Lai tạo bao gồm 2 bước: Lai ghép và Đột biến. Thông thường, để có thể tham gia quá trình Lai tạo, các NST sẽ phải trải qua bước Lựa chọn, trong đó GA sẽ dựa vào 1 thuật toán nào đó để chọn NST được Lai tạo. Kết quả sau quá trình Lai tạo là một quần thể NST con được sử dụng cho thế hệ NST tiếp theo.

2.2.1.Các nghiên cứu liên quan

- GA sẽ dựa vào thuật toán Thay thế để chọn các NST con và các NST từ quần thể hiện tại để đưa vào quần thể thế hệ mới. Bằng cách lặp lại quá trình này, GA có thể tìm được nhiều NST khác nhau với hi vọng là NST tốt sẽ xuất hiện ở quần thể thế hệ cuối. Thuật toán sẽ dừng lại sau khi số thế hệ vượt quá số lượng cho phép hoặc hàm fitness vượt qua ngưỡng mong muốn của bài toán.

2.2.1.Các nghiên cứu liên quan

Luồng hoạt động cơ bản của GA



2.2.2.Phương pháp đề xuất

2.2.2.1.Tổng quan về hệ thống

Ý tưởng chính đằng sau Hệ thống dịch máy dựa trên xác suất (SMT) đó là SMT sẽ tạo ra Bản dịch (T) từ Câu cần dịch (S), sao cho xác suất P(T|S) là tốt nhất có thể. Xác suất này có thể được tính dựa theo công thức Bayes như sau:

$$P(T|S) = P(S|T)*P(T)/P(S)$$

Trong đó: P(S|T): Xác suất câu T được dịch thành câu S

P(T): Xác suất xuất hiện của câu T

Để tối đa hóa xác suất P(T|S), ta sẽ cần tối đa hóa giá trị của P(S|T)*P(T). Giá trị của P(S) thể hiện xác suất mà ta nhận được câu S – thứ mà ta không thể tính toán được vì nó hoàn toàn phụ thuộc vào dữ liệu nhập vào của người dùng – vì vậy P(S) sẽ không tham gia vào quá trình tối đa hóa xác suất của ta.

2.2.2.Phương pháp đề xuất

2.2.2.1.Tổng quan về hệ thống

$$argmaxP(T|S)=argmax(P(S|T)*P(T))$$

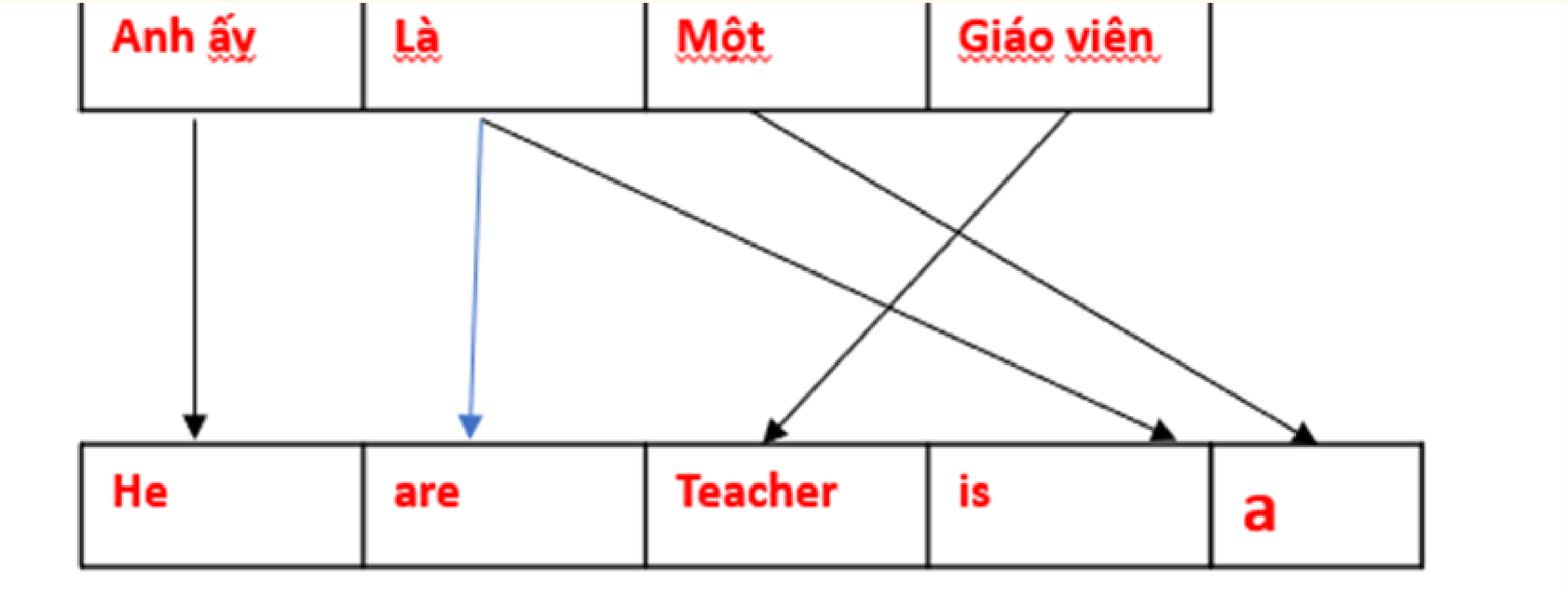
P(S|T) là xác suất nhận được câu nguồn S khi có Bản dịch T. Xác suất này là để giải quyết vấn đề về việc **lựa chọn các từ để dịch sao cho phù hợp**. P(S|T) sẽ được tính toán bởi Mô hình dịch thuật (Translation Model). P(T) là xác suất mà Bản dịch T là một bản dịch (câu) có nghĩa, hay nói cách khác là Bản dịch T có đúng **cấu trúc ngữ pháp** hay không. Xác suất này sẽ được tính bởi Mô hình Ngôn ngữ (Language Model). Ngoài 2 model kể trên, hệ thống SMT cần thêm thành phần thứ 3 được gọi là Decoder. Thành phần này đảm nhận việc tối đa hóa giá trị PS|T*P(T), và từ đó tìm được bản dịch với xác suất P(T|S) là lớn nhất. Với hệ thống của mình, chúng em lựa chọn Genetic Algorithm làm Decoder

2.2.2.2.Bài toán dịch câu tiếng anh sang tiếng việt

a.Translation Model (Mô hình Dịch thuật)

Mô hình dịch thuật dựa trên mối quan hệ song ngữ của hai câu từ ngữ liệu song song (parallel corpora).

Biểu thức P(S|T) có thể được hiểu là xác suất tạo ra câu Nguồn (S) từ Bản dịch (T), giá trị của nó có thể được ước tính từ một ngữ liệu song song (parallel corpora). Mỗi câu trong ngôn ngữ nguồn có một câu tương ứng trong ngôn ngữ đích. Ánh xạ (Alignment) là một tập hợp các mối quan hệ giữa các từ trong câu nguồn đến các từ trong câu đích với xác suất ước tính.



2.2.2.2.Bài toán dịch câu tiếng anh sang tiếng việt

b.Language Model (Mô hình Ngôn ngữ)

Mô hình ngôn ngữ thống kê là một phân phối xác suất trên các chuỗi từ. Với một câu có độ dài m, nó gán một xác suất P(w1,...,wm) cho toàn bộ câu.
Mô hình ngôn ngữ được sử dụng trong các ứng dụng xử lý ngôn ngữ tự nhiên khác nhau như Dịch máy, Nhận dạng giọng nói, Phân tích và truy xuất thông tin,...
Language Model được sử dụng để tìm đầu ra tốt nhất trong ngôn ngữ đích.
Language Model sử dụng phương trình sau để ước tính khả năng của một từ sẽ xuất hiện sau một chuỗi từ nhất định. Xác suất chuỗi từ đó được kí hiệu là P(w1 ... wm); dựa vào chain rule, ta có thể tính xác suất đó dựa theo công thức

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1})$$

trong đó m là số từ trong chuỗi (câu) và n là số lượng các từ đứng trước từ Wi

2.2.2.3.Hàm fitness

Để GA có thể được sử dụng làm Decoder cho SMT, hàm fitness của GA cần nhận vào 2 tham số, đó chính là các xác suất được tính bởi Language Model và Translation Model. Giá trị hàm fitness được tính dựa theo công thức sau:

$$fitness = a*\log(P(T|S)) + b*\log(P(T))$$

Trong đó: a: Tham số cho Translation Model

b: Tham số cho Language Model

Quá trình Lựa chọn

Áp dụng thuật toán Roulette wheel cho quá trình Lựa chọn. Theo phương thức này, xác suất mỗi NST sẽ được lựa chọn dựa trên giá trị fitness. Các NST với giá trị fitness cao sẽ có xác suất được lựa chọn cao hơn.

1. Tính tổng giá trị fitness của tất cả các NST trong quần thể - tổng S.
2. Sinh một số ngẫu nhiên nằm trong khoảng (0, S) – số r.
3. Lần lượt lấy từng NST trong quần thể và cộng dồn giá trị fitness của chúng từ 0 – tổng s. Khi s>r, dừng thuật toán và chọn NST vừa được lấy ra, đưa NST đó vào danh sách được Lai tạo.

2.2.2.3.Hàm fitness

Quá trình Lai tạo

Đối với bước Lai ghép, thuật toán Lai ghép đơn được sử dụng

1. Sinh ra NST con 1 có cùng số bộ gen với NST cha.
2. Sinh ra NST con 2 có cùng số bộ gen với NST mẹ.
3. for j=0 to length_of_selected_parent (Chọn NST có độ dài ngắn nhất)
    a. Chọn ngẫu nhiên NST cha hoặc mẹ
    b. Thêm từ ở vị trí j trong NST được chọn vào NST con 1
    c. Thêm từ ở vị trí j trong NST không được chọn vào NST con 2
4. for k=j to length_of_remaining_parent (NST không được chọn ở bước trước)
    a. Thêm từ ở vị trí k trong NST không được chọn ở bước trước vào NST con 1/2 (chọn NST con có độ dài lớn nhất)
5. Trả về 2 NST con.

2.2.2.3.Hàm fitness

Thuật toán Lai ghép đơn

Một số NST con sẽ được Đột biến. Hệ thống sẽ sinh ra một giá trị ngẫu nhiên p. Nếu p < P (P là một giá trị được cố định), quá trình Đột biến sẽ được thực hiện.

Quá trình Thay thế

Những NST cha mẹ sẽ được sắp xếp dựa trên giá trị hàm fitness. Những NST có giá trị fitness cao sẽ được giữ lại và đưa vào quần thể thế hệ tiếp theo.

2.2.3.Thực nghiệm và đánh giá kết quả

a.Mô tả tập dữ liệu

Tập dữ liệu bao gồm các câu đơn tiếng anh và tiếng việt. Có tất cả 200 câu bao gồm 100 câu tiếng việt và 100 tiếng anh. Mỗi câu gồm ít nhất 2 thành phần chính là chủ ngữ và vị ngữ, độ dài các câu là khác nhau.

b.Mô tả tập thực nghiệm

Bộ dữ liệu được chia thành hai bộ dữ liệu huấn luyện – kiểm tra với tỉ lệ 80%-20% để tiến hành quá trình thực nghiệm:

- Bộ dữ liệu thứ nhất là bộ dữ liệu huấn luyện (training dataset): bao gồm 80 câu tiếng việt và 80 câu tiếng anh dùng để huấn luyện mô hình.
- Bộ dữ liệu thứ hai là bộ dữ liệu kiểm tra (test dataset): bao gồm 20 tiếng việt và 20 câu tiếng anh dùng để kiểm tra mô hình.

c. Kết quả thực nghiệm

Với những câu có trong dữ liệu huấn luyện, hệ thống đưa ra kết quả chính xác khoảng 98%.

Với những câu nằm ngoài dữ liệu huấn luyện, độ chính xác của hệ thống đạt được khoảng 80%

2.2.4.Kết luận

Dịch ngôn ngữ là một trong những yếu tố quan trọng cho sự phát triển của xã hội ngày nay. Đề tài đã đưa ra phương pháp dịch câu tiếng việt sang tiếng anh có hiệu quả cao, bằng việc sử dụng giải thuật di truyền để tìm ra câu dịch tốt nhất so với ngôn ngữ gốc. Các kết quả thực nghiệm cũng cho thấy phương pháp của chúng em đủ nhanh, độ chính xác cao. Trong tương lai, chúng em sẽ cố gắng xây dựng một tập dữ liệu lớn hơn để hệ thống có thể dịch một cách chính xác hơn nữa. Nhược điểm của phương pháp hiện tại là độ chính xác khi dịch các từ không phổ biến chưa được cao. Chúng em sẽ tiếp tục cải tiến phương pháp và cung cấp giải pháp cho các vấn đề còn lại.

Chương 5: Thử nghiệm đánh giá chất lượng hệ thống

5.1.Tiêu chuẩn đánh giá

Hiện nay trong nước chưa ban hành đầy đủ các tiêu chuẩn, cũng như phương thức đánh giá chính quy nào về việc đánh giá tiêu chuẩn chất lượng một hệ thống dịch thuật AI. Vì vậy chúng tôi tham khảo các tiêu chuẩn cũng như phương pháp đánh giá của nước ngoài để áp dụng cho việc đánh giá hệ thống.

5.1.1.Đánh giá độ chính xác của hệ thống

- Phương pháp đánh giá là sử dụng phép đo BLEU trong dịch máy :

- BLUE là viết tắt của Bilingual Evaluation Understudy, là phương pháp đánh giá một bản dịch dựa trên các bản dịch tham khảo. BLEU được thiết kế để sử dụng trong dịch máy(Machine Translation), nhưng thực tế, phép đo này cũng được sử dụng trong các nhiệm vụ như tóm tắt văn bản, nhận dạng giọng nói, sinh nhãn ảnh v..v..
- Điều kiện tiên quyết để có thể sử dụng BLEU là bạn phải có một (hoặc nhiều) câu mẫu. Đối với bài toán dịch máy, câu mẫu chính là câu đầu ra của cặp câu trong tập dữ liệu.

5.1.1.Đánh giá độ chính xác của hệ thống

- Cách tính điểm : BLUE đánh giá một câu thông qua việc so khớp câu đó với các câu mẫu và cho thang điểm từ 0 (sai lệch tuyệt đối) đến 1 (khớp tuyệt đối).
- Trong Python, để tính điểm BLEU chúng ta có thể thực hiện dễ dàng với sự hỗ trợ của thư viện NLTK(Natural Language Toolkit).
- Hệ thống đạt chất lượng nếu điểm BLEU lớn hơn hoặc bằng 0.55.

5.1.2.Đánh giá sự ổn định của hệ thống

Phương pháp đánh giá :
- Yêu cầu hệ thống có thể chạy liên tục trong thời gian dài.
- Dịch câu và đưa ra kết quả trong thời gian ngắn.
- Trong quá trình chạy hệ thống bị lỗi thì phải tiến hành đánh giá lại từ đầu.

5.2.1 Đánh giá độ chính xác của hệ thống

Để đánh giá kết quả thực nghiệm, nhóm dự án tiến hành dịch 100 câu để lấy mẫu thử nghiệm, kết quả cụ thể như sau:

Precision = Độ chính xác của toàn hệ thống đạt 80% đạt yêu cầu.

5.2.2 Đánh giá độ ổn định của hệ thống

Kết quả đánh giá độ ổn định cũng cho thấy hệ thống có thể chạy liên tục, nhanh mà không xảy ra bất kỳ sự cố nào.

5.2.1 Đánh giá độ chính xác của hệ thống

Sau đây là một số hình ảnh và kết quả thu được khi chạy thử nghiệm trên thực tế.

Kết luận và Định hướng nghiên cứu trong tương lai

Hệ thống tự động dịch thuật là một hệ thống quan trọng và cần thiết giúp nâng cao năng lực ngoại ngữ, khả năng từ vựng góp phần xây dựng đô thị thông minh.

Kết quả thực nghiệm cũng cho thấy, thuật toán Giải thuật di truyền do nhóm nghiên cứu phát triển đạt độ chính xác cao, với thời gian xử lý nhanh hoàn toàn phù hợp với ứng dụng thời gian thực. Ngoài ra phần mềm do nhóm nghiên cứu phát triển cũng đạt độ ổn định cao đủ tiêu chuẩn để đưa vào hoạt động trên thực tế. Hệ thống của nhóm phát triển vẫn còn những khuyết điểm cần phải cải thiện thêm.

Kết luận và Định hướng nghiên cứu trong tương lai

Cơ sở dữ liệu vẫn chưa thực sự đầy đủ cần phải thu thập thêm để cải thiện độ chính xác huấn luyện mô hình. Tiếp theo nhóm nghiên cứu cũng có kế hoạch cải thiện kiến trúc mô hình nhằm tăng cường độ chính xác của mô hình hơn nữa. Ngoài ra hệ thống mới chỉ được thực nghiệm trong phạm vi của Học viện điều này chưa đảm bảo được hệ thống sẽ chạy tốt trong môi trường khác. Do vậy trong thời gian tới nhóm nghiên cứu sẽ tiến hành thực nghiệm đối với nhiều địa điểm môi trường khác nhau, để đảm bảo hệ thống có thể chạy tốt trong mọi điều kiện hoàn cảnh khác nhau.