|  |  |
| --- | --- |
|  | TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  **CHƯƠNG TRÌNH CHẤT LƯỢNG CAO** |

# ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP/ĐỒ ÁN TỐT NGHIỆP

**Tên đề tài:**

XÂY DỰNG BỘ NGỮ LIỆU GÁN NHÃN PASBIO

**Họ tên, chức danh GVHD: ThS.Tuấn Nguyên Hoài Đức**

**Sinh viên: Phạm Hữu Sang** (1653108) – **Hoàng Văn Thức** (1653139)

**Thể loại KLTN/ĐATN: Nghiên cứu lý thuyết và xây dựng ứng dụng minh họa**

**Thời gian thực hiện:** (20/01/2020 – 20/07/2020)

**Nội dung KLTN/ĐATN:**

|  |
| --- |
| * Nội dung: * Tìm hiểu về PAS và PASBio. * Tìm hiểu các phương pháp xây dựng Corpus. * Tìm hiểu phương pháp mở rộng Corpus bằng Virtual Example * Lựa chọn phương pháp phù hợp và thực hiện xây dựng bộ ngữ liệu gán nhãn PASBio * Thử nghiệm kết quả huấn luyện trên bộ ngữ liệu xây dựng được. * Yêu cầu: * Có kiến thức về máy học, khai khoáng dữ liệu. * Phương pháp thực hiện: (bỏ các chữ B1, B2… đi, chỉ gạch đầu dòng list ra thôi) * B1: Isolating : Tách bộ tập câu hiện có của PASbio thành 2 bộ : Seed và Archieve * B2: Làm giàu các câu trong 2 bộ trên ra nhiều cấu trúc ngữ pháp khác nhưng không thay đổi cách hành văn của câu nhưng giữ nguyên nội dung câu. * B3 : Chốt và quyết định cấu trúc XML của bộ ngữ liệu * B4 : Viết và test công cụ tạo internal virtual example (tool A) * B5 : Viết và test công cụ tạo external virtual example (tool B) * B6 : Dùng tool A phát sinh internal virtual example * B7 : Dùng tool B phát sinh external virtual example * B8 : Merge bộ Seed và bộ Archieve thành bộ ngữ liệu hợp nhất * Kết quả đạt được: * Báo cáo tổng hợp. * Bộ ngữ liệu gán nhãn PASBio |

**Các mốc thời gian nghiên cứu:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thời gian** | **Yêu cầu** | **Phân công** | |
| H.Sang | V.Thức |
| 29/03/2020 | Isolating:  Gọi toàn bộ tập câu hiện có của PASBio là bộ Seed.  Trong bộ Seed, những câu nào đồng thời xuất hiện 2 hoặc nhiều predicate của PASBio thì:   * Tách câu phức thành các câu đơn. Mỗi câu đơn chỉ có 1 predicate của PASBio. * Thêm những câu đơn này vào bộ Seed. * Cô lập câu phức gốc ra 1 bộ dữ liệu khác, chuyên dung để chứa những câu phức bị lọc ra, gọi là bộ Archieved.   Kết quả là ta có:   * Một bộ **Seed**, trong đó tất cả các câu đều chỉ có 1 predicate của PASBio   Một bộ **Archieve**, tromng đó chứa toàn những câu gốc mà có 2 predicate của PASBio trở lên. | Lập bộ Seed | Lập bộ Archieve |
| 19/04/2020 | Viết lại câu trong **Seed** và **Archieved**, mỗi câu ra càng nhiều cấu trúc ngữ pháp khác càng tốt (nhưng không thay đổi nội dung câu)  Sinh viên cần đề xuất với GVHD rằng sẽ phát sinh những biến thể ngữ pháp nào. Sau khi GVHD duyệt thì SV cứ như vậy mà áp dụng cho tất cả các câu trong PASBio | Làm giàu bộ Archieve | Làm giàu bộ Seed |
| 13/04/2020 | Chốt quyết định cấu trúc XML của bộ ngữ liệu. | Tìm hiểu và xây dựng cấu trúc XML cho bộ ngữ liệu | |
| 10/05/2020 | Viết và test ổn thỏa công cụ tạo ví dụ ảo internal (tool A): Nghĩa là công cụ hoán đổi các bộ đối số giữa các câu trong cùng một predicate để tạo ra các câu mới. | Viết tool | Test tool |
| 31/05/2020 | Viết và test ổn thỏa công cụ tạo ví dụ ảo external (tool B): Nghĩa là công cụ thay thế các bộ đối số do chuyên gia Công nghệ Sinh học cung cấp vào những câu đang có để tạo ra các câu mới. | Test tool | Viết tool |
| 14/06/2020 | Dùng tool A cho phát sinh ví dụ ảo internal. Kiểm tra kỹ lưỡng chất lượng của từng câu phát sinh ra có đúng ý đồ hay không. | Dùng tool phát sinh internal virtual example | Kiểm tra chất lượng kết quả |
| 28/06/2020 | Dùng tool B cho phát sinh ví dụ ảo external. Kiểm tra kỹ lưỡng chất lượng của từng câu phát sinh ra có đúng ý đồ hay không. | Kiểm tra chất lượng kết quả | Dùng tool phát sinh external virtual example |
| 12/07/2020 | Merge bộ Seed và bộ Archieve thành bộ ngữ liệu hợp nhất | Merge và kiểm tra lần cuối | |
| 08/03/2020 | Đọc paper | Đọc paper | Đọc paper |
| 22/03/2020 | Viết chương 1 của LV. Gửi GVHD review ngay khi viết xong. | Viết mục : - Lý do chọn đề tài  -Thách thức thách thức bài toán và hướng giải quyết đề tài | Viết mục :  -Giới thiệu đề tài  - Mục tiêu đề tài |
| 05/04/2020 | Viết chương 2 của LV. Gửi GVHD review ngay khi viết xong. | Viết mục :  -Cấu trúc đối số vị ngữ  -Cấu trúc đối số vị ngữ trong y sinh | Viết mục :  -Các khung đối số phổ biến  -Các khung đối số y sinh |
| 24/05/2020 | Viết chương 3 của LV. Gửi GVHD review ngay khi viết xong. | Minh chứng tính đúng đắn và phân tích lý do chọn hướng thực hiện | |
| 07/06/2020 | Viết chương 4 của LV. Gửi GVHD review ngay khi viết xong. | Mô tả cách xây dựng corpus của đề tài | |
| 21/06/2020 | Viết chương 5 của LV. Gửi GVHD review ngay khi viết xong. | Trình bày cách đánh giá tính đúng đắn của corpus | Xây dựng một mô hình học máy đơn giản để chứng minh kết quả |
| 05/07/2020 | Viết chương 6 của LV. Gửi GVHD review ngay khi viết xong. | Kết luận những gì được và chưa được của đề tài | Trình bày hướng phát triển |
| 12/07/2020 | Soạn slide báo cáo GVPB và HĐ | Soạn slide và báo cáo | |

|  |  |
| --- | --- |
| **Ý kiến của giảng viên hướng dẫn** | **TP. HCM, 01/03/2020** |
| ***Chữ ký của giảng viên hướng dẫn*** | ***Chữ ký (các) sinh viên*** |