ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thị Chăm

MÔ HÌNH HỌC MÁY SUỐT ĐỜI TRONG KHAI PHÁ VĂN BẢN Y SINH HỌC

Chuyên ngành: Hệ thống thông tin
Mã số: 9480104.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2022

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: PGS.TS Hà Quang Thụy

Phản biện: .....
.....
Phản biện: .....
.....
Phản biện: .....
.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại
vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại:
- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin – Thư viện, Đại học Quốc gia Hà Nội

MỞ ĐẦU

Học máy suốt đời (Lifelong Machine Learning, "học suốt đời", "học liên tục" (Continual Learning) hoặc "học không dừng" (Never-ending learning)), viết tắt là HMSĐ, là một kiểu học máy mới, liên tục tiếp nhận và thực hiện các bài toán học, lưu trữ lại các tri thức học được vào cơ sở tri thức (CSTT), chọn lựa tri thức đã được lưu lại trước đó để hỗ trợ giải quyết bài toán học mới xuất hiện. Chắt lọc tri thức (Knowledge Distillation/ Distilling Knowledge), viết tắt là CLTT, một kiểu học chuyển giao từ một mô hình phức tạp và cồng kềnh (mô hình giáo viên) đã được đào tạo trước tới một mô hình đơn giản và nhỏ hơn (mô hình sinh viên), là một giải pháp hữu dụng để lựa chọn tri thức quá khứ hữu dụng hỗ trợ bài toán hiện thời trong HMSĐ, trong đó CSTT đóng vai trò là mô hình giáo viên và bài toán học hiện tại đóng vai trò mô hình sinh viên. Số lượng công bố (bao gồm các luận án Tiến sỹ) về học giám sát suốt đời (học cơ bản, học sâu, học thế giới mở), học không giám sát suốt đời (học mô hình chủ đề suốt đời, học không giám sát khác), học tăng cường suốt đời, học khi làm việc, chắt lọc tri thức và áp dụng HMSĐ (trong xử lý ngôn ngữ tự nhiên, trong người máy, v.v.) tăng nhanh trong một vài năm gần đây. Luận án tập trung nghiên cứu, phát triển các kỹ thuật KD, chuyển tri thức hữu dụng từ CSTT của hệ thống HMSĐ phục vụ bài toán học hiện thời hiệu quả hơn.

Mục tiêu nghiên cứu của luận án là nghiên cứu và phát triển các kỹ thuật CLTT trong HMSĐ, tập trung vào học mô hình chủ đề suốt đời (CĐSĐ) và học sâu suốt đời đối với dữ liệu văn bản. Đề xuất các mô hình và thuật toán học mô hình CĐSĐ miền gần dựa trên CLTT từ các miền quá khứ gần và giải pháp CLTT mô hình học sâu quá khứ.

Đối tượng nghiên cứu của luận án là các kỹ thuật CLTT trong HMSĐ trên các miền dữ liệu văn bản: mô hình CĐSĐ, nhận dạng thực thể định danh dựa trên học sâu BiLSTM và áp dụng các mô hình này.

Phạm vi nghiên cứu của luận án được giới hạn trên các mô hình, kỹ thuật CLTT trong HMSĐ trên miền dữ liệu văn bản và áp dụng chúng.

Phương pháp nghiên cứu của luận án là vừa tiến hành phân tích định tính các khái niệm và mô hình để đề xuất các kỹ thuật CLTT phù hợp cho MHCĐ suốt đời, phân lớp NER suốt đời dựa trên học sâu BiLSTM vừa tiến hành các phân tích định lượng thông qua việc triển khai các mô hình thực nghiệm để kiểm chứng, đánh giá đối với các đề xuất của luận án.

Luận án có ba đóng góp chính (i) Đề xuất thuật toán và mô hình CĐSĐ miền gần CD-AMC dựa trên mô hình CĐSĐ AMC với giải pháp chắt lọc tri thức must-link và cannot-link chỉ từ các miền quá khứ gần thay vì từ tất cả các miền quá khứ [NTCham1]. Đề xuất hai cách thức xác định miền gần đối

với miền dữ liệu hiện tại (từ vựng–từ–chủ đề và dựa trên các bộ phân lớp văn bản quá khứ), áp dụng vào bài toán phân lớp đa nhãn tiếng Việt [NTCham1] và bài toán phân lớp quan điểm tiếng Anh [NTCham2], đồng thời, tiến hành đánh giá thực nghiệm các mô hình đề xuất. Hơn nữa, tiến hành kiểm định thống kê một mẫu theo *phân phối-t* (one-sample t test) về kỳ vọng quần thể giả thuyết khi chưa biết độ lệch chuẩn quần thể để minh chứng mô hình đề xuất thực sự có hiệu năng cao hơn so với AMC [NTCham1]; (ii) Đề xuất mô hình CĐSĐ miền gần hướng đích TCD-AMC kết hợp mô hình CĐSĐ miền gần CD-AMC với mô hình chủ đề hướng đích TTM và áp dụng vào bài toán phân lớp đa nhãn trích xuất khía cạnh trong khai phá quan điểm tiếng Việt [NTCham3]; (iii) Đề xuất mô hình HMSĐ chắt lọc tri thức tham số mô hình học sâu BiLSTM-KD-NER cho bài toán nhận dạng thực thể y sinh tiếng Việt và tiến hành thực nghiệm kiểm chứng, đánh giá đề xuất này [NTCham4].

CHƯƠNG 1. KHÁI QUÁT VỀ HỌC MÁY SUỐT ĐỜI, CHẮT LỌC TRI THỨC VÀ MÔ HÌNH CHỦ ĐỀ SUỐT ĐỜI

Chương này trình bày các kiến thức cơ bản về học máy suốt đời, chắt lọc tri thức, mô hình chủ đề suốt đời, tập dữ liệu thực nghiệm và các độ đo sử dụng để đánh giá hiệu năng của mô hình.

1.1 Học máy suốt đời

1.1.1 Sơ lược về lịch sử tiến hóa

HMSĐ được khởi đầu từ đầu thập niên 1990 (hệ thống CHILD của M. B. Ring, hệ thống EBNN của S. Thrun và T. M. Mitchell). Theo thời gian, các vấn đề về quản lý dòng bài toán, trích xuất tri thức và lưu trữ chúng vào cơ sở tri thức, lựa chọn tri thức quá khứ hữu ích để hỗ trợ việc giải quyết bài toán hiện tại nhằm nâng cao hiệu năng bài toán học hiện tại, v.v. trong HMSĐ ngày càng được làm sáng tỏ hơn, toàn diện hơn và sâu sắc hơn. Định nghĩa về hệ thống HMSĐ và khung hoạt động hệ thống này do Z. Chen và B. Liu (2016, 2018) đề nghị được coi là toàn diện nhất cho tới hiện nay về các thành phần và yếu tố của HMSĐ.

1.1.2 Định nghĩa và khung hệ thống học máy suốt đời

Học máy suốt đời (HMSĐ) là một quá trình học liên tục. Ở thời điểm bất kỳ, bộ học đã thực hiện một dãy N bài toán $T_1, T_2, \dots, T_N$ (được gọi là bài toán quá khứ) có các tập dữ liệu tương ứng là $D_1, D_2, \dots, D_N$. Các bài toán quá khứ có thể thuộc các kiểu khác nhau và từ các miền bài toán khác nhau. Khi xuất hiện bài toán mới T_{N+1} (được gọi là *bài toán hiện tại*) với tập dữ liệu D_{N+1} của nó, bộ học cần tận dụng tri thức quá khứ trong cơ sở tri thức (CSTT) S để giúp học bài toán T_{N+1}. Bài toán hiện tại có thể nhận được từ bên ngoài

hoặc do bộ học tự phát hiện. Mục đích của HMSĐ thường là làm tối ưu hiệu năng của bài toán hiện tại T_{N+1} song nó cũng có thể làm tối ưu hiệu năng của bất kỳ bài toán nào thuộc $\{T_1, T_2, \dots, T_N\}$ khi coi toàn bộ các bài toán còn lại (bao gồm T_{N+1}) như các bài toán quá khứ. HMSĐ có năm đặc điểm HMSĐ là (i) quá trình học liên tục, (ii) tích lũy và duy trì tri thức trong CSTT, (iii) sử dụng tri thức quá khứ đã tích lũy được để giúp việc học trong tương lai, (iv) phát hiện các bài toán mới, (v) học khi làm việc hoặc học theo công việc. Hệ thống HMSĐ có năm thành phần: Bộ quản lý bài toán, Bộ học tri thức, Bộ khai phá tri thức hướng bài toán, Cơ sở tri thức, Mô hình, Ứng dụng.

1.1.3 So sánh MHSĐ với các kiểu học máy truyền thống gần gũi

HMSĐ là khác biệt so với năm kiểu học máy truyền thống gần gũi là học chuyển đổi, học đa nhiệm/tác vụ theo lô, học trực tuyến đơn nhiệm, học tăng cường và siêu học.

1.1.4 Học thế giới mở và học khi làm việc

Học thế giới mở (open-world learning, nhận dạng/phân lớp thế giới mở hoặc TTNT thế giới mở) là bộ học cần nhận ra được những thứ chưa biết từ môi trường và học những thứ chưa biết đó để ngày càng thông minh hơn.

Học khi làm việc (Learning on the job) là kiểu HMSĐ mà việc khai thác tri thức được tiến hành không chỉ theo tri thức quá khứ có trong hệ thống mà còn qua tương tác với môi trường (bao gồm người sử dụng) để thu nhận dữ kiện, tri thức hỗ trợ việc học bài toán mới.

1.1.5 Hệ thống học ngôn ngữ không dừng NELL

Hệ thống học ngôn ngữ không dừng NELL (http://rtw.ml.cmu.edu/rtw/) là một hệ thống HMSĐ tiêu biểu hoạt động từ năm 2010 tới nay, có kiến trúc không ngừng phát triển cho phép các tác tử thông minh học được nhiều kiểu kiến thức, học liên tục tự giám sát trong nhiều năm, học tốt hơn theo thời gian để hình thành các tác vụ học mới với các trình diễn mới.

1.1.6 Thách thức đối với học máy suốt đời

Z. Chen và B. Liu (2018) chỉ ra thách thức: Tính đúng đắn của tri thức, Khả năng áp dụng tri thức, Biểu diễn và suy luận tri thức, Học với các bài toán thuộc nhiều loại và/hoặc từ các miền khác nhau, Học tự tạo động lực học, Học tự giám sát, Học ngôn ngữ tự nhiên suốt đời, Học theo thành phần.

1.2 Chắt lọc tri thức

Nén mô hình là việc tạo nên các mô hình nhỏ và đơn giản qua tận dụng và xấp xỉ được năng lực của các mô hình đồ sộ và phức tạp. Chắt lọc tri thức là một kiểu nén mô hình: một mô hình đào tạo đơn giản và nhỏ hơn (mô hình sinh viên) được thiết lập nhờ lấy mô hình phức tạp và cồng kềnh được đào tạo trước (mô hình giáo viên) làm mục tiêu học tập. Trong HMSĐ, CSTT là mô hình giáo viên và bài toán học hiện thời là mô hình sinh viên.

1.3 Mô hình chủ đề suốt đời

1.3.1 Mô hình chủ đề ẩn

Hai mô hình chủ đề truyền thống (pLDA và LDA) và một số phiên bản nâng cấp (MDK-LDA, GK-LDA, TTM (và BiTTM)) được giới thiệu.

1.3.2 Mô hình chủ đề suốt đời

Mô hình CĐSĐ là một kiểu HMSĐ không giám sát với bài toán học mô hình chủ đề. Hai kiểu tri thức thông dụng là must-link và cannot-link.

Thuật toán mô hình chủ đề suốt đời LTM

Thuật toán 1.1 Mô hình chủ đề suốt đời (LTM)

Đầu vào: Tập dữ liệu miền mới D_{N+1}; Cơ sở tri thức S
Đầu ra: Mô hình chủ đề miền mới A_{N+1}

```
A_{N+1} ← GibbsSampler(D_{N+1}, ∅, M) //Chạy lặp M lần không tri thức
For i = 1 to M do
    K_{N+1} ← TopicKnowledgeMiner(A_{N+1}, S)
    A_{N+1} ← GibbsSampler(D_{N+1}, K_{N+1}, 1) //Chạy 01 lần với tri thức
Endfor
S ← UpdateKB(A_{N+1}, S)
```

Thuật toán 1.2 mô tả hàm TopicKnowledgeMiner dùng CSTT S để hỗ trợ tinh chỉnh mô hình chủ đề A_{N+1}:

Thuật toán 1.2 TopicKnowledgeMiner

Đầu vào: Mô hình chủ đề miền mới A_{N+1}, Cơ sở tri thức S,
ngưỡng liên quan hai chủ đề $\pi > 0$.
Đầu ra: Tập tri thức must-link K_{N+1} được sử dụng để tinh chỉnh A_{N+1}

```
For mỗi p_chủ_đề s_k ∈ S do   // lặp với mọi chủ đề quá khứ
    j* = min_j KL-Devergence (a_j, s_k) //a_{j*} ∈ A_{N+1}: KL(a_{j*},s_k) nhỏ nhất
    If KL-Devergence(a_{j*}, s_k) ≤ π then
        M_{j*}^{N+1} ← M_{j*}^{N+1} ∪ {s_k} //Khởi thủy: M_{j*}^{N+1} rỗng.
    Endif
Endfor
K_{N+1} ← ∪_{j*} FIM (M_{j*}^{N+1})      // Tìm 2-mục phổ biến trong M_{j*}^{N+1}.
```

Thuật toán mô hình chủ đề suốt đời AMC

Thuật toán 1.3 Mô hình AMC

Đầu vào: Tập dữ liệu miền mới D_{N+1}, Cơ sở tri thức S
Đầu ra: Mô hình chủ đề miền mới A_{N+1}.

```
MK ← MustLinkMiner(S)
C = ∅ // C lưu trữ các cannot-link
A_{N+1} ← GibbsSampler(D_{N+1}, MK, C, M)
For r = 1 to R do
    C = C ∪ CannotLinkMiner(S, A_{N+1})
    A_{N+1} ← GibbsSampler(D_{N+1}, MK, C, N)
Endfor
S ← UpdateKB(A_{N+1}, S)
```

Thiếu sót của LTM và AMC

Tác giả của LTM và AMC, Z. Chen và B. Liu, chỉ ra thiếu sót về: (i) biểu diễn tri thứ khi (chỉ sử dụng hai kiểu tri thức must-link và cannot-link); (ii) khai thác tri thức; (iii) chuyển giao tri thức; (iv) lưu trữ và duy trì tri thức. Đặc biệt, giả định của LTM (và AMC) cho rằng mọi miền quá khứ đều liên quan và hữu ích cho miền hiện tại không phải lúc nào cũng đúng.

1.3.3 Bốn mô hình chủ đề suốt đời gần đây

Bốn mô hình CĐSĐ RLTM-SK (2017), JLTMMR (2019), LNTM (2020) và LCM (2021) được giới thiệu. Bốn mô hình này chưa xem xét giải quyết giả định của AMC (LTM) là tri thức từ mọi miền quá khứ luôn hữu ích.

1.4 Tập dữ liệu Hotels

Tập dữ liệu Hotels là tập các đánh giá tiếng Việt về miền khách sạn theo các khía cạnh (nhãn lớp) về Vị trí và giá cả, Nhân viên và Dịch vụ, Cơ sở vật chất, Phòng tiêu chuẩn và Đồ ăn. Tập dữ liệu có các đặc điểm sau:

- Số lượng ví dụ: $|Hotel|$=1493,
- Số lượng đặc trưng:$|Dim(Hotel)| = 1266$,
- Số lượng nhãn lớp có thể: $|L(Hotel)|$=5,
- Số nhãn trung bình của một ví dụ trong Hotel: $LCard(Hotel) = \frac{1}{|S|}\sum_{i=1}^{|Hotel|}|Y_i| = 1.250$,
- Mật độ nhãn khi chuẩn hóa nhãn số lượng nhãn có thể theo số nhãn trung bình của một ví dụ: $LDen(Hotel) = \frac{LCard(Hotel)}{L(Hotel)}$,
- Tập nhãn phân biệt khi đếm số lượng nhãn phân biệt trong Hotel: $DL(Hotel) = |\{Y: (x, Y) \in Hotel\}|$,
- Tỷ lệ nhãn phân biệt chuẩn hóa $DL(Hotel)$ theo số lượng ví dụ trong Hotel: $PDL(Hotel) = \frac{DL(Hotel)}{|Hotel|}$.

Tập dữ liệu Hotel được luận án sử dụng để tạo ra các tập dữ liệu D_{N+1} của bài toán mới T_{N+1}. Đặc điểm của tập dữ liệu Hotel cho dấu hiệu về mức độ khó khăn khi phân lớp đa nhãn đối với tập dữ liệu đó, chẳng hạn, mật độ nhãn càng thưa thì khó khăn càng lớn.

1.5 Các độ đo đánh giá hiệu năng phân lớp

Các độ đo hồi tưởng, chính xác và trung bình hài hòa F1 được giới thiệu.

1.6 Kết luận

Tổng hợp nội dung khảo sát trong Chương 1.

CHƯƠNG 2. MÔ HÌNH CHỦ ĐỀ SUỐT ĐỜI MIỀN GẦN

Chương này trình bày đề xuất về mô hình CĐSĐ miền gần và hai bài toán áp dụng của mô hình CĐSĐ miền gần.

2.1 Mô hình chủ đề suốt đời miền gần CD-AMC

2.1.1 Ý tưởng mô hình chủ đề miền gần

Luận án đề xuất ý tưởng chắt lọc tri thức mức cao (meta level) từ các miền quá khứ "gần" để hỗ trợ xây dựng mô hình chủ đề cho bài toán hiện tại.

2.1.2 Miền gần

Thuật toán AMC (và LTM) xác định một chủ đề quá khứ s_k là tương tự với một chủ đề a_{j^*} của bài toán hiện tại như sau: (i) tính độ đo phân kỳ Kullback-Leibler rời rạc KL(a_{j^*}, s_k) cho hai phân phối ứng với a_{j^*} và s_k, (ii) so sánh với một ngưỡng cho trước $\pi > 0$: nếu KL$(a_{j^*}, s_k) \leq \pi$ thì chủ đề s_k được coi là tương tự với chủ đề a_{j^*}.

Phân kỳ KL $D_{KL}(p(x), q(x))$ xác định một phân bố quan sát *q(x)* có thể thay thế cho một phân bố *p(x)* theo công thức sau:

$$D_{KL}(p(x), q(x)) = \sum_{x \in X} q(x) ln \frac{q(x)}{p(x)} \quad (2.1)$$

Khi tập giá trị quan sát X là rời rạc, công thức trên được viết lại thành:

$$D_{KL}(p(x), q(x)) = \sum_{i=1}^{n} q(x_i) ln \frac{q(x_i)}{p(x_i)} \quad (2.2)$$

Định nghĩa 2.1. Độ phân kỳ Kullback-Leibler (KL-Devergence) rời rạc thu gọm theo *k* được tính toán theo công thức:

$$D_{KL,k}(p(x), q(x)) = \sum_{i=1}^{k} q(x_i) ln \frac{q(x_i)}{p(x_i)} \quad (2.3)$$

Để đơn giản, luận án dùng độ đo tương tự cosin của hai túi từ với trọng số.

Định nghĩa 2.2. (Độ đo tương tự của hai túi từ với trọng số).

Cho hai túi từ có trọng số $A = \{(wa_i, pa_i)\}, B = \{(wb_i, pb_i)\}$, trong đó wa_i và wb_i là các từ, pa_i và pb_i là trọng số tương ứng của chúng.

Gọi C là từ vị gồm các từ ở cả A và B, tức là $C = \{wa_i\} \cup \{wb_i\}$. Gọi v_A, v_B là các véctơ trọng số (dựa trên tập từ vị C) của A và B, trong đó từ không có trong A hoặc B sẽ có trọng số bằng 0. Độ đo tương tự của A và B:

$$Similarity(A, B) = cosine(v_A, v_B) \quad (2.4)$$

Độ đo tương tự túi từ được dùng cho khái niệm chủ đề gần và miền gần.

Định nghĩa 2.3. (Chủ đề gần)

Gọi θ>0 là một ngưỡng cho trước, hai chủ đề khác nhau X và Y được gọi là gần nhau nếu:

$$similarity\ (Top_M(X), Top_M(Y)) \geq \theta \quad (2.5)$$

trong đó, $\mathrm{Top_M}(X)$ và $\mathrm{Top_M}(Y)$ là hai túi gồm M từ có trọng số cao nhất trong chủ đề X và Y.

Hàm $Similarity(\mathrm{Top_M}(X), \mathrm{Top_M}(Y))$ được gọi là độ đo chủ đề gần.

Định nghĩa 2.4. (Miền gần)

Gọi D_i và D_j là tập (miền) dữ liệu của hai bài toán T_i và T_j, tương ứng; V_i (V_j) là tập từ vị của D_i (D_j), $Topic_i$ ($Topic_j$) là mô hình chủ đề của D_i (D_j); miền D_j được gọi là gần miền D_i nếu thỏa mãn đồng thời các điều kiện sau:

1. Mức từ vựng:

$$\frac{|V_i \cap V_j|}{|V_i|} + \frac{|V_i \cap V_j|}{|V_j|} \geq \theta_1 \qquad (2.6)$$

2. Mức từ tốp đầu trong từ vựng:

$$Similary(Top_M(V_i), Top_M(V_j)) \geq \theta_2 \qquad (2.7)$$

3. Mức chủ đề:

Số lượng chủ đề tương tự của T_j với T_i đảm bảo điều kiện:

$$\frac{|\{t_2 \in Topics(T_j) | t_1 \in Topics(T_i) \wedge similary(t_2, t_1) \geq \theta_3\}|}{|Topics(T_j)|} \geq \theta_4 \qquad (2.8)$$

trong đó, θ_1, θ_2, θ_3, và θ_4 là các ngưỡng có giá trị dương.

Luận án xây dựng mô hình CĐSĐ miền gần dựa trên chắt lọc tri thức miền gần là chỉ sử dụng các mẫu tri thức must-link và cannot-link từ các miền quá khứ gần thay vì từ mọi miền quá khứ như AMC. Khi cài đặt, các điều kiện (2.6), (2.7), (2.8) được kiểm tra tuần tự, nếu một điều kiện vi phạm một cách nghiêm trọng thì không cần kiểm tra các điều kiện phía sau. Hơn nữa, để nhận được tập miền gần khác rỗng, các ngưỡng θ_1 - θ_4 (ưu tiên điều chỉnh hai ngưỡng θ_3, θ_4 vì liên quan tới mức chủ đề) được điều chỉnh.

2.1.3 Thuật toán mô hình chủ đề suốt đời miền gần CD-AMC

Luận án đề xuất thuật toán mô hình CĐSĐ miền gần ***CD-AMC*** (*Close Domain - AMC*) được phát triển từ thuật toán AMC:

Thuật toán 2.1 CD-AMC;

Đầu vào: Tập dữ liệu miền mới D_{N+1}, Cơ sở tri thức S

Đầu ra: Mô hình chủ đề của miền mới A_{N+1}, Cơ sở tri thức S được cập nhật

Nội dung

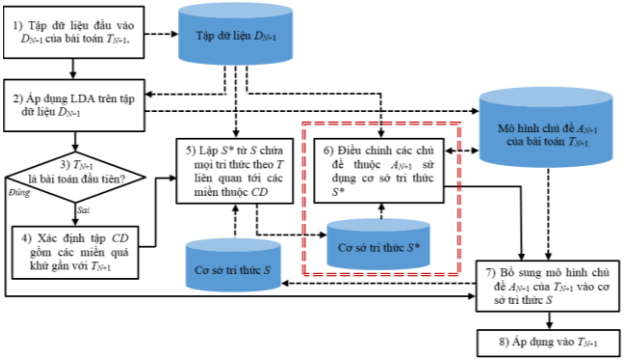
```
1  A_N ← LDA (D_{N+1})                    // Xây dựng mô hình chủ đề A_{N+1}
2  CD ← Close (A_{N+1}, T_{N+1})          // Xác định tập CD các miền quá khứ gần T_{N+1}
3  S* ← Reduce (S, CD)                    // S* là thu hẹp của S theo CM
4  ML ← MustLinkMiner (S*)                // Tìm mọi tri thức must-link từ S*
5  C = ∅                                  // C lưu trữ các cannot-link
6  A_{N+1} ← GibbsSampler(D_{N+1}, ML, C, M)  // Lặp M lần Gibbs với tập tri thức must-link
                                              ML và chưa dùng cannot-link
7  For r = 1 to R do
8        C = C ∪ CannotLinkMiner(S*, A_{N+1}) //Tìm các cannot-link
9        A_{N+1} ← GibbsSampler(D_{N+1}, ML, C, N)
10 Endfor
11 S ← UpdateKB(A_{N+1}, S)
```

Thuật toán CD-AMC có một số thay đổi so với Thuật toán 1.3 AMC như sau: (i) Bổ sung ba dòng lệnh 1-3 để xây dựng mô hình *LDA* cho T_{N+1}, tìm tập *CD* gồm mọi miền quá khứ gần với bài toán T_{N+1}, lập S^* là tập thu gọn

của S bao gồm mọi dữ liệu, thông tin và tri thức từ tập CD, (ii) các dòng lệnh 4-11 là nội dung thuật toán **Thuật toán 1.3** (AMC) song được áp dụng đối với S^* thay vì S.

2.1.4 Khung hệ thống mô hình chủ đề suốt đời miền gần

Hình 2.2 mô tả các bước thực hiện trong mô hình CĐSĐ miền gần dựa trên thuật toán CD-AMC.



Hình 2.2 Khung mô hình chủ đề suốt đời miền gần CD-AMC của luận án

2.1.5 Phần mềm thực thi mô hình chủ đề suốt đời miền gần

Luận án sử dụng mã nguồn Java tại https://github.com/czyuan/AMC với các hướng dẫn cụ thể và các tập dữ liệu thực thi mô hình AMC. Trong các trường hợp áp dụng mô hình CĐSĐ miền gần, luận án bổ sung mô-đun xác định miền gần vào mã nguồn AMC.

2.2 Mô hình chủ đề suốt đời miền gần cho phân lớp đa nhãn văn bản tiếng Việt

Áp dụng mô hình chủ đề CD-AMC vào phân lớp đa nhãn văn bản tiếng Việt: bộ dữ liệu Hotels là nguồn tạo ra các tập dữ liệu D_{N+1} của bài toán hiện thời và ba tập dữ liệu khác là các tập dữ liệu của các bài toán quá khứ.

2.2.1 Phát biểu bài toán

Cho trước:

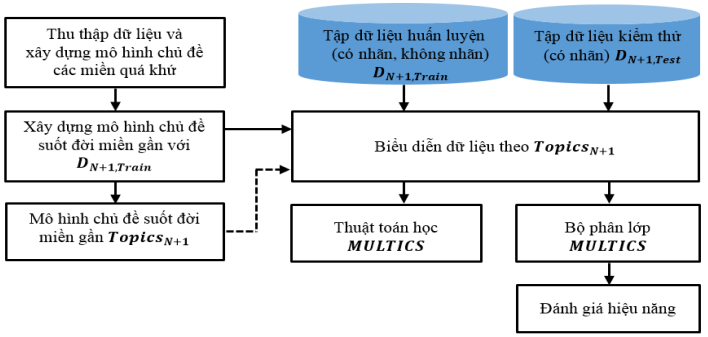
- D_{N+1}: tập dữ liệu văn bản gồm các nhận xét của người dùng về miền khách sạn (Hotels) với tập nhãn L gồm năm nhãn.
- Tập dữ liệu D_{N+1} được chia thành hai tập, tập dữ liệu huấn luyện (gồm dữ liệu có nhãn và không có nhãn) ký hiệu là $D_{N+1,Train}$ và tập $D_{N+1,test}$ làm tập dữ liệu đánh giá (chỉ bao gồm các dữ liệu có nhãn).

Tìm:

Bộ phân lớp đa nhãn $Multics$ có sử dụng mô hình CĐSĐ miền gần DC-AMC để biểu diễn dữ liệu.

2.2.2 Mô hình giải quyết

Hình 2.3 mô tả quá trình áp dụng mô hình CĐSĐ miền gần DC-AMC vào bài toán phân lớp đa nhãn.



Hình 2.3 Mô hình áp dụng CD-AMC vào phân lớp đa nhãn của luận án

2.2.3 Thực nghiệm và nhận xét

2.2.3.1 Các tập dữ liệu

Bảng 2.1 giới thiệu các tập dữ liệu được luận án sử dụng để đánh giá hiệu năng mô hình phân lớp gồm (i) ba tập dữ liệu quá khứ (D_1, D_2, D_3); (ii) D_4 được sử dụng để tạo ra các phiên bản khác nhau cho bài toán hiện tại. D_4 được chia thành hai phần: tập dữ liệu test gồm 300 đánh giá ($D_{4,test}$), và 1,493 đánh giá để thiết lập năm bộ dữ liệu huấn luyện nhỏ cho bài toán hiện tại là D_{4a}, D_{4b}, D_{4c}, D_{4d} và D_{4e} có kích thước tương ứng là 100, 200, 400, 600 và 1000 đánh giá.

Bảng 2.1 Các tập dữ liệu của các miền khác nhau

Tập dữ liệu	Số lượng đánh giá	Tên miền
D_1	26,800	Du lịch và Khách sạn *(Tourism and Hotels)*
D_2	8,093	Nhà hàng *(Restaurants)*
D_3	1,441	Điện thoại di động *(Mobile phones)*
D_4	1,493	Khách sạn *(Hotels)*

2.2.3.2 Kịch bản thực nghiệm

Với mỗi tập dữ liệu hiện tại D_{N+1} từ $\{D_{4a}, D_{4b}, D_{4c}, D_{4d}, D_{4e}\}$, luận án tiến hành cài đặt ba thực nghiệm sau:

Thực nghiệm 1: Thực hiện phân lớp sử dụng tập đặc trưng theo mô hình chủ đề ẩn LDA của tập dữ liệu đầu vào để xây dựng tập chủ đề $Topics_{N+1}$ của bài toán hiện tại, đánh giá hiệu năng của phương pháp học cô lập.

Thực nghiệm 2: Thực hiện phân lớp sử dụng tập đặc trưng theo mô hình chủ đề của bài toán hiện tại $Topics_{N+1}$ được xây dựng bằng mô hình CĐSĐ AMC, đánh giá hiệu quả của mô hình CĐSĐ AMC cho tập dữ liệu đầu vào có kích thước nhỏ.

Thực nghiệm 3: Thực hiện phân lớp sử dụng tập đặc trưng theo mô hình chủ đề $Topics_{N+1}$ được xây dựng bằng mô hình CĐSĐ miền gần CD-AMC, đánh giá mức độ ảnh hưởng của (i) tri thức từ các miền gần và (ii) tập dữ liệu đầu vào có kích thước nhỏ. Các tham số tương ứng như sau: θ_1 = 0.5, θ_2 = 0.9, θ_3 = 0.1; θ_4 = 0.2, M = 20.

2.2.3.3 Kết quả thực nghiệm và nhận xét

Trong bước đầu tiên của thực nghiệm, luận án xác định các miền quá khứ gần với miền dữ liệu của bài toán hiện tại. Kết quả tính toán cho thấy chỉ có miền D_1 được xác định là gần với các miền D_{4a}, D_{4b}, D_{4c}, D_{4d}, D_{4e}. Do đó, chỉ sử dụng D_1 để tinh chỉnh mô hình chủ đề ẩn trên mỗi tập dữ liệu D_{N+1} hiện tại trong {D_{4a}, D_{4b}, D_{4c}, D_{4d}, D_{4e}}.

Ba độ đo hiệu năng được sử dụng để đánh giá hiệu năng bộ phân lớp là độ chính xác P, độ hồi tưởng R và độ đo hài hòa F1.

Bảng 2.3 Kết quả của các kịch bản thực nghiệm của LDA, AMC, CD-AMC (P là độ chính xác, R là độ hồi tưởng, F là độ đo hài hòa)

Số chủ đề	Tập dữ liệu huấn luyện	Mô hình LDA			Mô hình AMC			Mô hình CD-AMC		
		P%	R%	F%	P%	R%	F%	P%	R%	F%
10	D4a	63.26	50.38	56.09	65.19	52.42	58.11	62.94	54.72	58.54
	D4b	70.68	55.24	62.01	74.68	58.52	65.62	75.32	59.03	66.19
	D4c	80.72	67.43	73.48	82.50	67.18	74.05	83.07	67.43	74.44
	D4d	82.42	68.96	75.09	84.38	68.7	75.74	85.05	69.47	76.47
	D4e	82.31	71.50	76.53	83.58	71.25	76.92	83.28	72.99	77.80
15	D4a	62.94	50.13	55.81	63.14	52.72	57.46	62.18	53.64	57.59
	D4b	71.13	54.76	61.88	73.70	57.76	64.76	73.70	58.36	65.14
	D4c	84.01	68.19	75.28	84.01	68.19	75.28	84.01	69.19	75.89
	D4d	84.47	69.21	76.08	84.74	69.21	76.19	84.74	69.21	76.19
	D4e	82.26	72.11	76.85	83.58	71.25	76.92	84.52	72.26	77.91
20	D4a	62.94	50.13	55.81	63.14	51.27	56.59	62.50	52.62	57.14
	D4b	72.70	55.76	63.11	73.70	57.76	64.76	74.03	58.02	65.05
	D4c	84.01	68.19	75.28	84.01	68.19	75.28	84.01	68.19	75.28
	D4d	84.47	69.21	76.08	84.74	69.21	76.19	84.78	69.47	76.36
	D4e	84.23	72.01	77.64	83.58	73.68	78.32	84.82	73.52	78.77
25	D4a	61.18	49.17	54.53	63.14	50.13	55.89	62.50	51.62	56.54
	D4b	72.43	54.70	62.33	73.70	57.76	64.76	73.38	57.51	64.48
	D4c	83.93	67.12	74.59	84.01	68.19	75.28	84.01	68.19	75.28
	D4d	84.35	69.42	76.16	84.74	69.21	76.19	84.78	69.47	76.36
	D4e	83.93	71.52	77.23	83.58	71.25	76.92	84.82	72.52	78.19

Các kết quả thực nghiệm cho thấy: trong hầu hết các kịch bản, tất cả các hệ thống thường nhận được kết quả tốt hơn khi kích thước tập dữ liệu huấn luyện tăng từ 100 (D_{4a}) lên 1000 (D_{4e}) đánh giá. Mô hình CĐSĐ miền gần CD-AMC cho kết quả tốt hơn so với mô hình AMC trong tất cả các nhóm thực nghiệm với mức độ cải thiện khoảng 1% ở mọi trường hợp. Sự thay đổi của hiệu năng cho thấy kích thước của tập dữ liệu hiện tại ảnh hưởng đến hiệu năng phân lớp, tức là mô hình CĐSĐ miền gần thực hiện tốt hơn ngay cả khi tập dữ liệu huấn luyện có số lượng đánh giá nhỏ hơn. Mô hình CĐSĐ miền gần CD-AMC đạt kết quả tốt nhất là 78.77% với 20 chủ đề.

2.2.4 Kiểm định hiệu năng CD-AMC so với LDA và AMC

Luận án tiến hành kiểm định theo *phân phối-t* một mẫu (one-sample t test) với 19 bậc tự do về giả thuyết trung bình mẫu quần thể khi chưa biết độ lệch chuẩn quần thể để chứng minh hiệu năng của mô hình CD-AMC thực sự cao hơn mô hình AMC là không thể bị bác bỏ.

2.3 Mô hình chủ đề suốt đời miền gần theo bộ phân lớp quá khứ

2.3.1 Mô hình chủ đề suốt đời miền gần CCD-AMC

Mục này trình bày mô hình CĐSĐ miền gần dựa trên các bộ phân lớp quá khứ CCD-AMC, đây là giải pháp xác định bộ phân lớp quá khứ liên quan tới tập dữ liệu D_{N+1} của bài toán hiện tại nhằm xác định miền gần vào mô hình CĐSĐ miền gần như sau:

Cho một tác vụ PLNPSĐ với N bài toán phân lớp nhị phân đã học trong quá khứ T_i với D_i, $Topics_i$, m_i lần lượt là tập dữ liệu, mô hình chủ đề và bộ phân lớp nhị phân của mỗi T_i.

Cho tập dữ liệu D_{N+1} của bài toán T_{N+1}, cần xây dựng mô hình chủ đề cho tập dữ liệu D_{N+1} để biểu diễn dữ liệu trong xây dựng bộ phân lớp nhị phân m_{N+1}.

Quá trình hoạt động của mô hình CĐSĐ CCD-AMC tương tự như CD-AMC với một khác biệt duy nhất là cách thức xác định miền gần dựa theo hai định nghĩa 2.5 và 2.6.

Định nghĩa 2.5. (Miền dữ liệu quá khứ gần của một điểm dữ liệu hiện tại)
Điểm dữ liệu hiện tại $x \in D_{N+1}$ được gọi là gần với tập dữ liệu quá khứ D_i khi và chỉ khi x được đoán nhận theo m_i, nghĩa là x là một ví dụ dương theo m_i (ví dụ dương của bài toán quá khứ T_i).

Định nghĩa 2.6. (Miền quá khứ gần với miền hiện tại)
Tập dữ liệu quá khứ D_i được gọi là gần với tập dữ liệu hiện tại D_{N+1} khi và chỉ khi:

$$\frac{|x \in D_{N+1} \wedge m_i(x) \text{là dương}|}{|D_{N+1}|} \geq \theta_{PL} \qquad (2.9)$$

trong đó $\theta_{PL} > 0$ là một ngưỡng cho trước.

Miền quá khứ tương ứng được gọi là gần với miền dữ liệu hiện tại.

Tập các bộ tài nguyên (D_i, $Topics_i$, m_i) của các miền quá khứ gần với miền hiện tại lập thành cơ sở tri thức miền gần trong xây dựng mô hình CĐSĐ miền gần của bài toán hiện tại.

Thuật toán CD-AMC và mô hình CD-AMC không thay đổi nội dung trong trường hợp mô hình CĐSĐ CCD-AMC vì chỉ có sự thay đổi việc xác định miền gần.

2.3.2 Áp dụng vào bài toán phân lớp quan điểm dựa trên học sâu

Luận án áp dụng mô hình CĐSĐ miền gần CCD-AMC vào bài toán phân lớp quan điểm.

2.3.2.1 Các tập dữ liệu

Luận án sử dụng các tập dữ liệu trong một nghiên cứu của Z. Chen và cộng sự (2018) gồm 20 tập dữ liệu tương ứng với 20 miền sản phẩm, mỗi miền có 1.000 bài đánh giá được thu tập từ trang web Amazon.com.

Nhằm kiểm chứng hiệu năng của mô hình CCD-AMC với một lượng nhỏ dữ liệu được gắn nhãn trong miền hiện tại, luận án không sử dụng toàn bộ 1.000 đánh giá cho miền hiện tại. Thay vào đó, luận án tạo ra 5 bộ dữ liệu miền hiện tại khác nhau với số lượng đánh giá lần lượt là 20, 40, 60, 80 và 100 để huấn luyện mô hình và chỉ một tập dữ liệu kiểm thử (D_{test}) có 100 đánh giá được dùng trong cả năm kịch bản; năm tập dữ liệu huấn luyện và tập dữ liệu kiểm thử tuân thủ tỷ lệ đánh giá dương gấp bốn lần đánh giá âm.

2.3.2.2 Kịch bản thực nghiệm

Với mỗi tập dữ liệu huấn luyện hiện tại, ba kịch bản sau được tiến hành:

(i) Xây dựng tập đặc trưng theo mô hình LDA từ tập dữ liệu hiện tại (không dùng tri thức trước), đây là phương án cơ sở, ký hiệu là LDA.

(ii) Xây dựng tập đặc trưng theo mô hình CĐSĐ AMC sử dụng toàn bộ 19 miền còn lại là miền quá khứ, làm phương án cơ sở đối sánh, ký hiệu là AMC;

(iii) Xây dựng tập đặc trưng theo mô hình CĐSĐ miền gần CCD-AMC, ký hiệu là CCD-AMC.

Độ đo hài hòa F được sử dụng để đánh giá hiệu năng phân lớp quan điểm.

2.3.2.3 Kết quả thực nghiệm và nhận xét

Mỗi kịch bản cấu hình số lượng chủ đề khác nhau cho mô hình LDA, AMC và CCD-AMC lần lượt là 10, 15, 25 và sử dụng các thuật toán phân lớp Decision tree, k-nearest neighbors, MultiLayer Perceptrons và Naïve Bayes. Kết quả cho thấy hai mô hình CĐSĐ AMC và CCD-AMC thực hiện tốt hơn mô hình cơ sở LDA trong hầu hết thực nghiệm. Trong nhiều trường hợp, CCD-AMC đã cải thiện đáng kể hiệu năng so với AMC, chứng tỏ rằng chắt lọc tri thức lựa chọn miền gần cung cấp các đóng góp có ý nghĩa hơn cho bước phân lớp. Các miền gần có thể đã loại bỏ được nhiễu, tức là dữ liệu từ các miền quá khứ khác không liên quan đến miền hiện tại, do đó, chúng giúp cải thiện hiệu năng của hệ thống phân lớp. Vì các bài đánh giá về sản phẩm thường là những câu ngắn và luận án sử dụng tập dữ liệu miền hiện tại khá nhỏ (từ 20 đến 100), cho nên số lượng chủ đề không nên quá nhỏ hoặc quá lớn. Thuật toán mạng nơ-ron đa tầng MLP và Thuật toán Bay-et "ngây thơ" (Naïve Bayes) sử dụng mô hình CĐSĐ miền gần CCD-AMC mang lại hiệu năng tốt nhất khi số lượng chủ đề bằng 15. Khi kích thước mô hình chủ đề là 25, trong nhiều trường hợp, hiệu năng phân lớp thấp hơn so với kích thước

mô hình chủ đề là 15 đối với cả ba mô hình LDA, AMC và CCD-AMC, lý do có thể là các đánh giá là văn bản ngắn.

2.4 Kết luận

Mục này tổng kết các kết quả nghiên cứu ở Chương 2.

CHƯƠNG 3. MÔ HÌNH CHỦ ĐỀ SUỐT ĐỜI MIỀN GẦN HƯỚNG ĐÍCH

Chương này đề xuất mô hình CĐSĐ TCD-AMC, là sự kết hợp của mô hình CĐSĐ miền gần CD-AMC với mô hình chủ đề hướng đích TTM (Targeted Topic Modeling) do S. Wang và cộng sự đề xuất năm 2016 và áp dụng TCD-AMC vào phân lớp đa nhãn văn bản tiếng Việt dựa trên học sâu.

3.1 Mô hình chủ đề hướng đích

Mô hình chủ đề hướng đích (S.Wang và cộng sự, 2016 và 2018), một biến thể hướng tri thức của LDA, trong đó "đích" là một từ/cụm từ do người dùng đưa vào để dẫn dắt việc tạo sinh các chủ đề "ẩn" cho mô hình chủ đề.

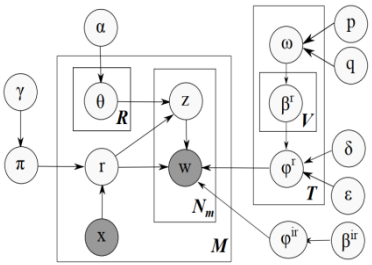
3.1.1 Bài toán

Cho trước:

C là một kho ngữ liệu lớn gồm các tài liệu về một miền dữ liệu văn bản rộng lớn nào đó.

S là một tập các từ khóa "đích" biểu diễn khía cạnh được người dùng quan tâm (khía cạnh đích).

Tìm: Mô hình chủ đề *TTM* thực sự đáp ứng "đích" S.



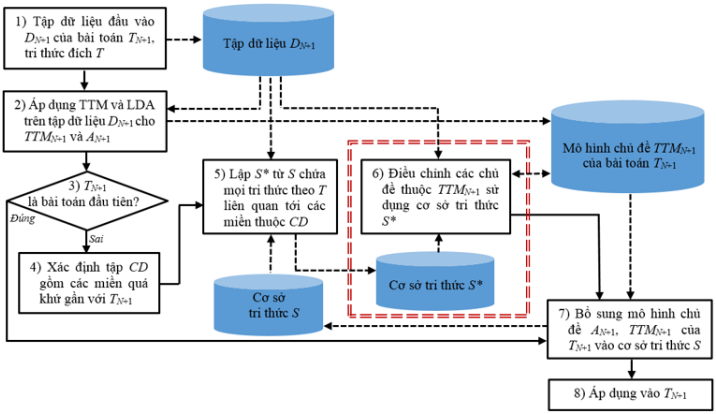
Hình 3.1 Mô hình TTM

3.1.2 Mô hình TTM

Với giả thiết mỗi tài liệu chỉ đề cập tới một khía cạnh, khi khía cạnh đích được chỉ dẫn thì một tài liệu cụ thể chỉ có thể là liên quan hoặc không liên quan tới khía cạnh đó; mô hình TTM bổ sung biến trạng thái r cùng với một vài kết nối tương ứng vào mô hình chủ đề LDA (sơ đồ hoạt động của mô hình TTM được mô tả ở Hình 3.1).

3.2 Đề xuất mô hình chủ đề suốt đời miền gần hướng đích TCD-AMC

Luận án đề xuất mô hình CĐSĐ miền gần hướng đích TCD-AMC kết hợp mô hình CĐSĐ CD-AMC và mô hình TTM như được mô tả trên Hình 3.2.



Hình 3.2 Sơ đồ xây dựng mô hình CĐSĐ miền gần hướng đích

Quá trình 8 bước xây dựng mô hình CĐSĐ TCD-AMC như sau:

- **Bước 1**. Nhập tập dữ liệu D_{N+1} của bài toán mới T_{N+1}, tri thức T,
- **Bước 2**. Áp dụng mô hình hóa chủ đề TTM và LDA cho D_{N+1} nhận được mô hình chủ đề TTM_{N+1} và A_{N+1} xuất phát,
- **Bước 3**. Nếu T_{N+1} là bài toán đầu tiên vào hệ thống thì chuyển tới Bước 7 (Tương tự CD-AMC),
- **Bước 4**. Xác định tập M bao gồm mọi miền quá khứ gần với T_{N+1}. Như đã được đề cập, để đảm bảo tập các miền quá khứ gần M khác rỗng, các ngưỡng xác định miền gần được điều chỉnh theo các giá trị tính toán được tương ứng. (Tương tự CD-AMC),
- **Bước 5**. Lập cơ sở tri thức S^* từ S chứa mọi dữ liệu, thông tin, tri thức liên quan tới các miền quá khứ thuộc M (Tương tự CD-AMC),
- **Bước 6**. Điều chỉnh mô hình chủ đề TTM_{N+1} (tương tự như thực hiện trong AMC) dựa trên cơ sở tri thức S^* theo đích T,
- **Bước 7**. Bổ sung tập dữ liệu D_{N+1}, mô hình chủ đề TTM_{N+1} và mô hình chủ đề A_{N+1} của bài toán T_{N+1} vào cơ sở tri thức S để sử dụng cho các bài toán tiếp theo.
- **Bước 8**. Sử dụng mô hình CĐSĐ miền gần hướng đích TTM_{N+1} vào việc giải quyết bài toán T_{N+1}.

Bước 6 điều chỉnh mô hình chủ đề TTM_{N+1} theo đích T có một thay đổi là sử dụng thuật toán FIM khai phá mẫu phổ biến (hoặc mẫu âm) có độ dài 2 với ràng buộc đích T thay vì thuật toán FIM khai phá mẫu phổ biến (hoặc mẫu âm) tổng quát.

3.3 Mô hình chủ đề suốt đời miền gần hướng đích cho phân lớp đa nhãn văn bản tiếng Việt

Mục này trình bày áp dụng mô hình CDSĐ miền gần hướng đích vào bài toán xây dựng bộ phân lớp đa nhãn dựa trên học sâu vừa tận dụng ưu điểm của độ đo miền gần để làm giàu tri thức là tìm và sử dụng các mẫu tri thức có ích từ các bài toán đã học trong quá khứ, vừa kết hợp với mô hình chủ đề

hướng đích để khai phá chi tiết hơn từng khía cạnh được gán nhãn của một ý kiến (hay quan điểm) của người dùng nhằm cải tiến chất lượng mô hình phân lớp đa nhãn.

3.3.1 Phát biểu bài toán

Cho trước:

D là tập tài liệu đầu vào đã được gán nhãn với một tập nhãn L gồm q nhãn, tức là $L = \{l_1, l_2, \dots, l_q\}$, trong đó mỗi tài liệu trong D được gán một tập con không rỗng của tập nhãn, tức là $label(d) \subseteq L, label(d) \neq \emptyset, \forall d \in D$. D là tập dữ liệu huấn luyện, cũng là tập dữ liệu của bài toán hiện tại.

Một đích T là một từ (hoặc cụm từ) do người sử dụng đưa vào để dẫn dắt việc xây dựng mô hình chủ đề (như được đề cập với *TTM*).

Một "vũ trụ dữ liệu không nhãn" gồm k tập dữ liệu văn bản không nhãn $\{D_i\}_{i=1,k}$, trong đó mỗi miền dữ liệu D_i có thể có liên quan hoặc không liên quan tới D.

M là tập từ nhúng được huấn luyện trước có sẵn và công khai cho ngôn ngữ chứa miền dữ liệu D và đích T.

Yêu cầu: Xây dựng bộ phân lớp học sâu đa nhãn văn bản dựa trên mô hình CĐSĐ miền gần hướng đích T đối với tập dữ liệu đầu vào D.

3.3.2 Mô hình giải quyết

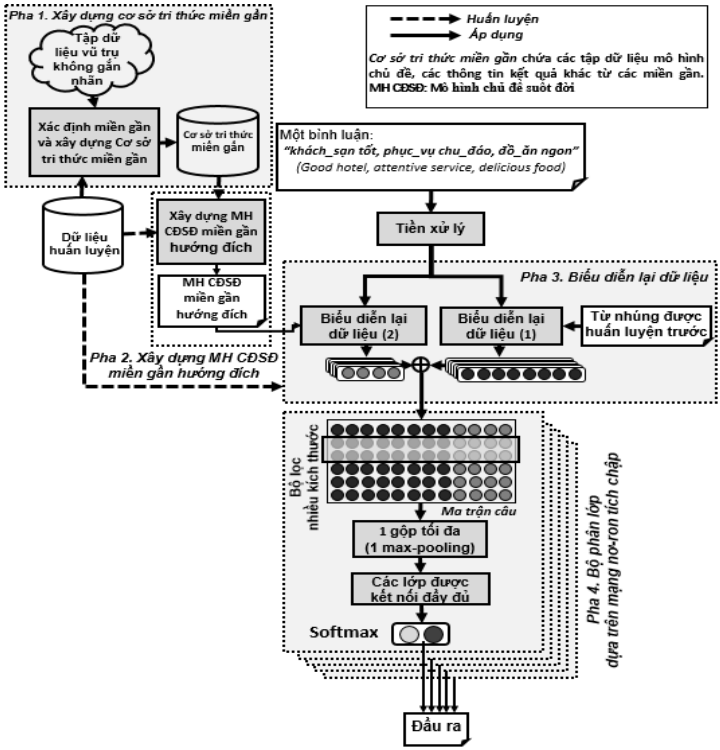
Luận án xây dựng mô hình phân lớp học sâu đa nhãn dựa trên mô hình CĐSĐ miền gần hướng đích gồm 4 pha (mô tả trong Hình 3.3) như sau:

Pha 1: Xây dựng cơ sở tri thức miền gần đối với tập dữ liệu D của bài toán. Đầu tiên, luận án áp dụng độ đo miền gần trên mỗi cặp tập dữ liệu (D_i, D) trên ba mức: mức từ tốp đầu, mức chủ đề và mức miền để đánh giá sự tương tự của hai tập dữ liệu nhằm tìm các tập dữ liệu gần với tập D. Kết quả, nhận được một *tập dữ liệu miền gần* sau khi bỏ qua các tập dữ liệu không/ít có khả năng cung cấp tri thức hữu ích trong việc nâng cao hiệu năng xây dựng mô hình chủ đề đối với D. Tiếp theo, xây dựng mô hình chủ đề đối với toàn bộ miền gần D_i. Sau đó, lập Cơ sở tri thức miền gần bao gồm các tập dữ liệu, các mô hình chủ đề và các thông tin khác của các tập dữ liệu miền gần D_i.

Pha 2: Xây dựng mô hình CĐSĐ miền gần hướng đích dựa trên tập dữ liệu huấn luyện của D và cơ sở tri thức miền gần.

Pha 3: Xây dựng biểu diễn dữ liệu. Sau khi tiền xử lý, dữ liệu bài toán được làm giàu với hai loại thông tin: (i) các từ nhúng được huấn luyện trước bằng FastText gồm véctơ 300 chiều, biểu diễn các từ dưới dạng tổng của véctơ bỏ qua gram (thông tin ngữ cảnh bên ngoài) và các véctơ n-gram và (ii) Biểu diễn dựa trên mô hình TTM gồm véctơ 15 chiều *n-hot* các từ đầu vào xuất

hiện trong danh sách từ khóa của các chủ đề được tạo bởi mô hình TTM. Việc nối hai véctơ này là một biểu diễn của mỗi từ đầu vào (véctơ 315 chiều).



Hình 3.3 Khung học sâu phân lớp đa nhãn dựa trên mô hình CĐSĐ miền gần hướng đích

Pha 4: Áp dụng mạng nơ-ron sâu để xây dựng bộ phân lớp đa nhãn

Mạng nơ-ron sâu với dữ liệu đầu vào là các tài liệu được biểu diễn dưới dạng ma trận 315×30 tạo thành bộ phân lớp đa nhãn, phân lớp từng tài liệu mới cho từng nhãn. Mạng bao gồm: (i) một lớp đầu vào là một tập các ma trận có kích thước 315×30 đại diện cho các câu từ bước trước đó, (ii) một lớp tích hợp, (iii) một lớp tổng hợp tối đa (*max-pooling*), và (iv) hai lớp được kết nối đầy đủ với lớp soft-max tạo thành đầu ra là các khía cạnh có thể có cho mỗi câu.

3.3.3 Thực nghiệm và nhận xét

3.3.3.1 Tập dữ liệu

Luận án sử dụng các tập dữ liệu như Chương 2 gồm: i) D_1 gồm các nhận xét về miền Du lịch và Khách sạn *(Tourism and Hotels)*; ii) D_2 gồm các nhận xét về Nhà hàng *(Restaurants)*; ii) D_3 là miền Điện thoại di động *(Mobile phones)*.

Các tập dữ liệu D_1, D_2, D_3 lần lượt là tập dữ liệu của ba bài toán quá khứ T_1, T_2, T_3. D là tập dữ liệu của bài toán hiện tại, gồm các tài liệu tiếng Việt đánh giá về Khách sạn có tên là *Hotels2* (nâng cấp từ tập dữ liệu Hotel).

Bảng 3.3 Các khía cạnh và quan điểm của các đánh giá trong tập dữ liệu

		Nhân viên và Dịch vụ	Tiện nghi, tiêu chuẩn phòng	Đồ ăn	Vị trí và giá cả	Cơ sở vật chất
Số lượng đánh giá	**Tích cực**	**640**	**569**	**403**	**376**	**298**
Độ dài các đánh giá (đếm các từ)	Dài nhất	45	41	45	45	45
	Ngắn nhất	2	2	2	2	2
	Trung bình	9.2	10.1	9.7	9.7	9.5
Số lượng đánh giá	**Tiêu cực**	**853**	**924**	**1,090**	**1,117**	**1,195**
Độ dài các đánh giá (đếm các từ)	Dài nhất	41	45	41	41	37
	Ngắn nhất	2	2	2	2	2
	Trung bình	8.1	7.6	8.2	7.6	7.1

Để áp dụng thuật toán BR (Binary Relevance) trên tập dữ liệu D, luận án đã chuyển tập dữ liệu đa nhãn D thành tập dữ liệu đơn nhãn, trong đó mỗi đánh giá chỉ được liên kết với một nhãn (một khía cạnh). Bảng 3.3 mô tả chi tiết về tập dữ liệu hiện tại sau khi được phân chia theo từng khía cạnh, đồng thời cho biết số lượng nhận xét tích cực/tiêu cực, độ dài lớn nhất/nhỏ nhất/trung bình cho từng khía cạnh cụ thể.

3.3.3.2 Kịch bản thực nghiệm

Để cho thấy hiệu quả của mô hình đề xuất, luận án đã thiết kế sáu kịch bản thực nghiệm, mỗi kịch bản bao gồm toàn bộ hoặc loại trừ một thành phần trong khung đề xuất nhằm đánh hiệu quả đóng góp của từng thành phần.

Thực nghiệm 1: gồm đầy đủ bốn pha của mô hình đề xuất,

Thực nghiệm 2: Tập dữ liệu D chỉ được biểu diễn dựa trên các từ nhúng được huấn luyện trước; sử dụng mô hình phân lớp CNN.

Thực nghiệm 3: chỉ sử dụng mô hình TTM trên tập dữ liệu D mà không sử dụng tập dữ liệu gần, phân lớp bằng CNN.

Thực nghiệm 4: thay bộ phân lớp CNN bằng một số bộ phân lớp phổ biến khác như *Random Forest*, *Support Vector Machine*, *k Nearest Neighbor*.

Thực nghiệm 5: chỉ sử dụng các đặc trưng của mô hình TTM làm đầu vào của mô hình phân lớp CNN.

Thực nghiệm 6: không sử dụng tri thức từ tập dữ liệu gần để huấn luyện mô hình CNN.

Độ đo được sử dụng: quá trình thực nghiệm cho thấy, do lấy trung bình vĩ mô với trọng số bằng nhau cho mỗi nhãn lớp, trong khi lấy trung bình vi mô cho trọng số bằng nhau cho mỗi tài liệu phân lớp. Do đó, luận án lấy trung bình vi mô gần hơn nhiều so với các lớp thống trị để đánh giá hiệu năng bộ phân lớp.

3.3.3.4 Kết quả thực nghiệm và nhận xét

Bảng 3.6 mô tả kết quả của các kịch bản thực nghiệm (tỷ lệ %).
Số in đậm là giá trị lớn nhất, số in nghiêng là giá trị lớn thứ hai trong mỗi cột.

Kịch bản thực nghiệm	Trung bình vi mô (%)			Thay đổi F
	P	R	F	
Thực nghiệm 1 (Toàn bộ hệ thống)	89.27	**76.53**	**82.41**	-
Thực nghiệm 2 (Không sử dụng TTM)	88.09	*75.62*	81.38	-1.03
Thực nghiệm 3 (Mô hình TTM không sử dụng miền gần)	89.28	75.52	*81.83*	-0.58
Thực nghiệm 4 (Thay CNN bằng Random Forest)	87.15	73.58	79.79	-2.62
Thực nghiệm 4 (Thay CNN bằng SVM)	**92.48**	52.12	66.67	-15.74
Thực nghiệm 4 (Thay CNN bằng KNN)	87.50	62.74	73.08	-9.33
Thực nghiệm 5 (CNN không sử dụng từ nhúng)	84.16	70.12	76.50	-5.91
Thực nghiệm 6 (CNN không sử dụng miền gần)	90.56	71.68	80.02	-2.39

Bảng 3.6 cho thấy kết quả của các kịch bản thực nghiệm ở trên nhằm đánh giá sự đóng góp của từng thành phần trong mô hình đề xuất. Kết quả đạt cao nhất ở thực nghiệm 1 thể hiện hiệu quả của mô hình đề xuất với bốn thành phần. Trong thực nghiệm 2 hệ thống thực nghiệm được cài đặt mà không sử dụng mô hình TTM và thực nghiệm 3 sử dụng mô hình TTM mà không sử dụng tập dữ liệu từ các miền gần cho TTM. Hiệu năng của bộ phân lớp thấp hơn một chút so với hiệu năng của hệ thống với đầy đủ các thành phần. Đây có thể là do số lượng chủ đề nhỏ (15 chủ đề) so với kích thước của véctơ từ nhúng (300 chiều). Tuy nhiên, những kết quả này vẫn cho thấy hiệu quả của mô hình TTM đối với mô hình đề xuất.

Kết quả trên nhóm Thực nghiệm 4 cho thấy mô hình CNN hoạt động tốt hơn các bộ phân lớp còn lại. Kết quả trong Thực nghiệm 5, Thực nghiệm 6 minh chứng cho sự đóng góp của từng thành phần trong việc huấn luyện mô hình CNN. Bên cạnh đó, luận án cũng so sánh kết quả của mô hình đề xuất với các mô hình nghiên cứu khác đã thực hiện trên cùng một tập dữ liệu đánh giá về Hotels như ở Chương 2 (xem Bảng 3.7).

Mô hình "*" (Pham và cộng sự, 2017) bao gồm một số thành phần biểu diễn TFIDF với các đặc trưng chủ đề ẩn của mô hình LDA và các đặc trưng thông tin tương quan (*Mutual Information*: MI) cho bộ phân lớp bán giám sát MULTICS; mô hình "+"(Pham và cộng sự, 2017a) được xây dựng từ bộ phân lớp bán giám sát MASS. Mô hình phân lớp đa nhãn dựa trên học sâu và mô hình chủ đề suốt đời miền gần hướng đích thực hiện tốt hơn mô hình CĐSĐ CD-AMC (Chương 2) và mô hình "*" mà không sử dụng dữ liệu không có nhãn. Tuy nhiên, hiệu năng mô hình phân lớp đa nhãn dựa trên học sâu và mô hình chủ đề suốt đời miền gần hướng đích ở đây thấp hơn một chút so với các mô hình sử dụng bộ phân lớp bán giám sát (được đánh dấu màu xám), có thể có lý do từ các dữ liệu

không có nhãn trong bộ phân lớp bán giám sát (mô hình "*"). Như vậy, mô hình phân lớp đa nhãn học sâu dựa trên mô hình CĐSĐ miền gần hướng đích không chỉ tận dụng được tri thức có ích của các miền gần mà còn tập trung hơn đối với từng khía cạnh được chỉ định.

Bảng 3.7 Kết quả của một số mô hình khác trên cùng tập dữ liệu thực nghiệm

Mô hình khác	Trung bình vi mô (%)		
	Precision	***Recall***	***F***
Mô hình đề xuất	89.27	76.53	82.41
MULTICS kết hợp mô hình CĐSĐ miền gần (Chương 2)	84.52	72.26	77.91
MULTICS (*)	80.10	79.60	79.80
MULTICS + phân lớp bán giám sát (*)	82.40	83.90	83.20
MULTICS+ TFIDF + LDA + MI (*)	N/A	N/A	80.60
MULTICS+ TFIDF + LDA + MI + phân lớp bán giám sát (*)	N/A	N/A	83.90
MASS (phân lớp bán giám sát) (+)	81.60	83.30	82.40

3.4 Kết luận

Mục này tổng kết các kết quả nghiên cứu ở Chương 3.

CHƯƠNG 4. CHẮT LỌC TRI THỨC HỌC SÂU SUỐT ĐỜI VÀ ÁP DỤNG VÀO NHẬN DẠNG THỰC THỂ Y SINH TIẾNG VIỆT

Chương này trình bày đề xuất kỹ thuật chắt lọc tri thức mô hình học sâu suốt đời và áp dụng vào nhận dạng thực thể y sinh tiếng Việt.

4.1 Nhận dạng thực thể định danh dựa trên chắt lọc tri thức và học máy suốt đời

Việc giới thiệu sơ bộ về Mô hình chắt lọc tri thức cho nhận dạng thực thể định danh MTM-STM (T. Mehmood và cộng sự, 2020), Mô hình chắt lọc tri thức đa hạt nhận dạng thực thể định danh (X. Zhou và cộng sự, 2021), Mô hình học liên lục nhận dạng thực thể định danh (N. Monaikul và cộng sự, 2021), mô hình DeepLML-NER nhận dạng thực thể định danh tiếng Việt (N. V. Nguyen và công sự, 2019) nhằm chỉ ra rằng mô hình HMSĐ BiLSTM-KD-NER do luận án đề xuất là có các điểm mới khác biệt.

4.2. Mô hình HMSĐ BiLSTM-KD-NER chắt lọc tri thức học sâu nhận dạng thực thể y sinh tiếng Việt

4.2.1 Phát biểu bài toán

Đầu vào:

- D: tập dữ liệu văn bản y sinh. Tập dữ liệu D được chia thành hai tập con: tập dữ liệu thứ nhất làm các tập dữ liệu quá khứ và tập dữ liệu thứ hai làm tập dữ liệu hiện tại (huấn luyện và đánh giá).
- L: tập 7 nhãn nhận dạng thực thể y sinh cho trước.

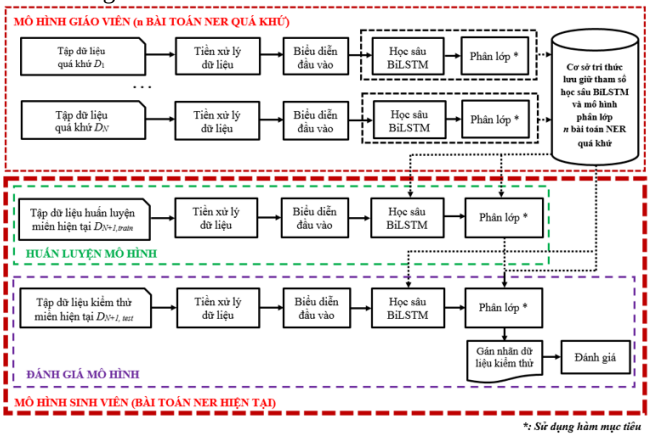
- M: một mô hình nhúng từ đã được luấn luyện trước.
- S: cơ sở tri thức.
- d: tài liệu cần gán nhãn thực thể.

Đầu ra:

- Các câu thuộc tài liệu d đã được gán nhãn chuỗi.
- Cơ sở tri thức S được cập nhật sau khi tài liệu d được gán nhãn thực thể.

4.2.2 Mô hình chung

Luận án đề xuất mô hình HMSĐ BiLSTM-KD-NER chắt lọc tri thức cho bài toán nhận dạng thực thể y sinh dạng mô hình giáo viên – mô hình sinh viên được mô tả trong Hình 4.5.



Hình 4.5 Sơ đồ kiến trúc BiLSTM-KD-NER chắt lọc mô hình học sâu suốt đời quá khứ cho nhận dạng thực thể định danh

Mô hình giáo viên: thực hiện việc chắt lọc tri thức từ n tập dữ liệu quá khứ để chuyển cho mô hình sinh viên. Luận án tập trung chắt lọc tri thức từ hai thành phần của mô hình NER là (i) tham số mô hình Bi-LSTM và (ii) mô hình phân lớp. Với mỗi tập dữ liệu quá khứ, mô hình NER đầu-cuối được huấn luyện để có được các tri thức mô hình nói trên;

Mỗi tập dữ liệu đầu vào được tiền xử lý qua bốn bước: phân đoạn câu, mã hóa, phân đoạn từ tiếng Việt và gắn thẻ POS. Tiếp đó, dữ liệu được đưa vào bước biểu diễn đầu vào để chuyển đổi từng dãy từ thành ma trận đầu vào cho học sâu BiLSTM. Luận án dùng một mạng nơ-ron BiLSTM để nhận các tham số mô hình. Dãy đầu ra của mô hình Bi-LSTM được sử dụng để xây dựng mô hình phân lớp gán thẻ NER. Bộ tham số Bi-LSTM và mô hình phân lớp của mỗi bài toán NER quá khứ được lưu trữ vào cơ sở tri thức để hỗ trợ xây dựng mô hình phân lớp đối với bài toán NER hiện tại.

Mô hình sinh viên gồm hai pha: huấn luyện mô hình và đánh giá mô hình.

Pha huấn luyện mô hình, mô hình NER cho bài toán hiện tại được xây dựng dựa trên tập dữ liệu huấn luyện miền hiện tại. Hai bước chuẩn bị đầu vào

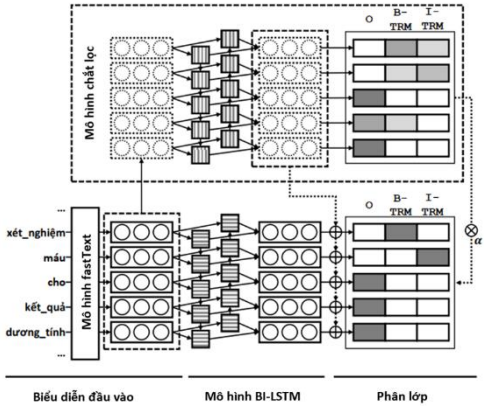
(tiền xử lý và biểu diễn đầu vào) tương tự như các bước tương ứng trong mô hình giáo viên. Mô hình sinh viên cũng sử dụng học sâu BiLSTM và xây dựng mô hình phân lớp NER như mô hình giáo viên, tuy nhiên, tri thức chắt lọc từ cơ sở tri thức của hệ thống được khai thác để nâng cao hiệu năng của học sâu BiLSTM và xây dựng mô hình phân lớp NER.

Pha đánh giá mô hình, mô hình NER của bài toán hiện tại được đánh giá theo các độ đo độ chính xác (P), độ hồi tưởng (R) và trung bình F.

4.2.3 Chi tiết giải pháp chắt lọc tri thức trong mô hình giáo viên

Hình 4.6 mô tả kiến trúc chắt lọc tri thức học sâu suốt đời nhận dạng thực thể định danh. Mô hình học sâu NER gồm ba pha chính (Biểu diễn đầu vào, Mô hình hóa Bi-LSTM, Phân lớp thực thể) và được chia thành hai kênh có kiến trúc tương tự là *Kênh mô hình chính* và *Kênh chắt lọc mô hình*.

Với một tài liệu đầu vào đã qua tiền xử lý, mô hình fastText được sử dụng với đặc trưng nhúng từ được huấn luyện trước và mô hình biểu diễn thẻ POS để tạo ra ma trận đầu vào. Hai mạng Bi-LSTM song hành để mô hình hóa thông tin dãy. Cuối cùng, hai mô hình perceptron đa tầng được kết hợp để gán nhãn thẻ NER tương ứng cho một từ vị.



Hình 4.6 Kiến trúc chắt lọc tri thức học sâu suốt đời nhận dạng thực thể định danh

Biểu diễn đầu vào: Ở pha biểu diễn đầu vào, mỗi token được chuyển thành một vectơ $x_e \in \mathbb{R}^d$, trong đó *d* là kích thước nhúng. Sử dụng ba loại thông tin: các nhúng được huấn luyện trước bằng fastText, các nhúng dựa trên ký tự và nhúng thẻ POS. Cuối cùng, các nhúng từ, nhúng ký tự và nhúng thẻ POS được kết hợp lại và chuyển đổi chúng thành nhúng từ vị cuối cùng.

Mô hình Bi-LSTM: Luận án xây dựng một mạng nơ ron hồi quy (RNN) với bộ nhớ dài ngắn hạn (LSTM) trên ma trận đầu vào để tính toán trạng thái ẩn h_t cho mỗi từ vị x_t tại điểm t.

Phân lớp: Đối với pha phân lớp, vectơ ẩn của mỗi từ vị từ tầng trước đó được đưa vào mạng perceptron đa tầng kết nối đầy đủ (multi-layer perceptron network: MLP).

Bảng 4.3 Phân tích hiệu năng của hệ thống

Mô hình	DISEASE			CHEMICAL			TREATMENT			Trung bình vĩ mô		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CRF	85.04	73.23	78.69	73.94	77.21	75.54	76.07	76.07	76.07	78.35	75.50	76.77
RNN	89.03	75.30	81.59	81.89	79.14	80.49	82.67	79.33	80.97	84.53	77.93	81.02
LSTM	87.07	83.29	85.14	86.51	80.50	83.40	84.55	80.11	82.27	86.04	81.30	83.60
BiLSTM	87.46	87.99	87.72	90.78	79.60	84.82	84.52	78.63	81.47	87.59	82.07	84.67
+ w dữ liệu tăng cường*	88.16	88.28	88.22	91.37	80.20	85.42	84.48	79.28	81.80	88.00	82.59	85.15
BiLSTM+CRF	88.78	89.17	88.97	86.19	85.51	85.82	84.07	82.29	83.04	86.34	85.65	85.94
+ w dữ liệu tăng cường*	89.66	90.55	90.11	86.66	85.17	85.91	84.30	82.23	83.25	86.87	85.98	86.42
BiLSTM-KD-NER	90.68	**91.21**	**90.95**	95.27	82.41	88.38	87.95	**82.19**	**84.97**	91.30	**85.27**	**88.10**
+ w CRF	**91.27**	90.17	90.71	**95.32**	**82.83**	**88.64**	**88.10**	81.36	84.59	**91.56**	84.78	87.98
-w/o chắt lọc BiLSTM	88.92	89.18	89.05	93.59	79.55	86.00	86.39	79.69	82.90	89.63	82.81	85.98
-w/o dự đoán chắt lọc	88.36	89.29	88.82	92.98	80.53	86.31	85.58	80.37	82.89	88.97	83.39	86.01

4.2.4 Thực nghiệm và đánh giá BiLSTM-KD-NER

Tập dữ liệu y sinh tiếng Việt dùng cho nhận dạng thực thể gồm (i) tập dữ liệu gồm 84 bài trực tuyến làm tập dữ liệu quá khứ, và (ii) tập gồm 82 báo cáo khoa học làm tập dữ liệu hiện tại (dữ liệu huấn luyện và kiểm thử).

Hàm mục tiêu: Luận án sử dụng hàm mục tiêu phạt entropy chéo (công thức 4.12) để huấn luyện mô hình học sâu và các mô hình chắt lọc như sau:

$$L(\theta) = -\sum_{i=0}^{K} \hat{y}_i \log \hat{y}_i + \lambda \|\theta\|^2 \qquad (4.12)$$

trong đó, $\hat{y} \in \{0,1\}^{(K+1)}$ biểu thị vectơ one-hot đại diện cho nhãn mục tiêu và λ là hệ số quy tắc hóa. Xác thực chéo 5 lần trên tập dữ liệu hiện tại. Quá trình huấn luyện và kiểm thử được tiến hành 10 lần với các mồi (seed) ngẫu nhiên khác nhau và sử dụng trung bình các kết quả của các lần thực hiện.

Thực nghiệm: tiến hành 3 kịch bản sau đây:

Thực nghiệm 1: Đánh giá hiệu năng hệ thống với các kỹ thuật khác nhau.

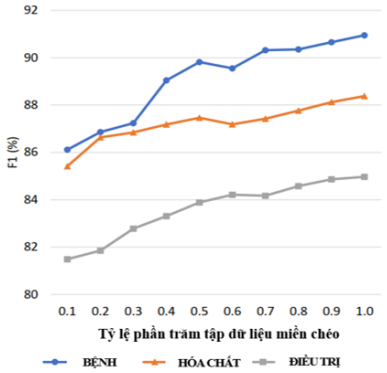
Thực nghiệm 2: So sánh kết quả của mô hình đề xuất khi thay đổi kích thước của tập dữ liệu quá khứ.

Thực nghiệm 3: So sánh kết quả của mô hình thông qua sự thay đổi ngưỡng quyết định của mô hình phân lớp được chắt lọc.

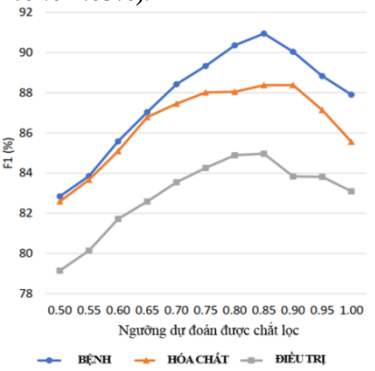
Kết quả thực nghiệm và nhận xét:

Thực nghiệm 1: so sánh mô hình đề xuất BiLSTM-KD-NER với ba mô hình học máy phổ biến gồm Trường ngẫu nhiên có điều điện, Mạng nơ-ron hồi quy và mạng bộ nhớ dài ngắn hạn và hai mô hình học sâu điển hình cho bài toán NER là BiLSTM và BiLSTM-CRF. Kết quả so sánh (Bảng 4.3) cho thấy mô hình BiLSTM-CRF đạt kết quả cao nhất trên tập dữ liệu hiện tại với kết quả trung bình vĩ mô là 85.94%, mô hình CRF có khả năng sửa lỗi các dự đoán của mô hình BiLSTM với mức cải thiện đạt 1.27%. Dữ liệu tăng cường từ tin tức trực tuyến mang lại sự cải thiện đáng kể cho cả mô hình BiLSTM và mô hình BiLSTM-CRF với sự thay đổi của F1 là 0.48%.

Mô hình chắt lọc tri thức học sâu suốt đời cho kết quả (theo F1) là vượt trội so với các mô hình khác trên các thực thể Bệnh, Hóa chất/thuốc và Điều trị F1 lần lượt cao hơn 3.71%, 3.20% và 3.43% so với mô hình BiLSTM cơ sở. Kết quả thực nghiệm cho thấy việc chắt lọc cả BiLSTM và mô hình phân lớp đều có tác động tích cực (tương ứng là 2.12% và 2.09%).



Hình 4.9 So sánh F1 dựa trên sự thay đổi kích thước của tập dữ liệu quá khứ



Hình 4.10 So sánh F1 dựa trên sự thay đổi ngưỡng quyết định của mô hình phân lớp BiLSTM-KD-NER

Thực nghiệm 2: thay đổi kích thước của tập dữ liệu quá khứ với tỷ lệ sử dụng từ 10% đến 100%. Hình 4.9 cho thấy F1 trên ba loại thực thể tỷ lệ thuận với tỷ lệ phần trăm dùng tập dữ liệu tin trực tuyến để chắt lọc tri thức.
Thực nghiệm 3: tăng dần ngưỡng quyết định trong khoảng [0, 1] thì kết quả cũng tăng theo tới khi ngưỡng đạt 0.85 và giảm xuống xấp xỉ bằng mô hình cơ sở khi cho ngưỡng quyết định tăng từ 0.85 lên 1.0.

4.3 Kết luận

Mục này tổng kết các kết quả nghiên cứu ở Chương 4.

KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU TIẾP THEO

Luận án có ba đóng góp chính sau đây:

- Đề xuất thuật toán và mô hình CĐSĐ miền gần CD-AMC dựa trên mô hình CĐSĐ AMC với giải pháp chắt lọc tri thức must-link và cannot-link chỉ từ các miền quá khứ gần. Một khung chung áp dụng mô hình CĐSĐ miền gần vào các bài toán phân tích văn bản được đề xuất. Hai cách thức xác định miền gần với miền dữ liệu hiện tại được đề nghị. Triển khai áp dụng mô hình CĐSĐ miền gần vào bài toán phân lớp đa nhãn tiếng Việt và bài toán phân lớp quan điểm tiếng Anh. Thực hiện kiểm định thống kê một mẫu theo *phân phối-t* (one-sample t test) về kỳ vọng quần thể giả thuyết khi chưa biết độ

lệch chuẩn quần thể để minh chứng mô hình đề xuất thực sự có hiệu năng cao hơn so với AMC

- Đề xuất mô hình CĐSĐ miền gần hướng đích TCD-AMC kết hợp CD-AMC với mô hình chủ đề hướng đích TTM. Triển khai áp dụng mô hình CĐSĐ miền gần hướng đích TCD-AMC vào bài toán phân lớp đa nhãn văn bản tiếng Việt dựa trên học sâu. Kết quả thực nghiệm trên sáu phương án cài đặt khác nhau cho thấy hiệu quả của khung đề xuất TCD-AMC so với mô hình AMC và các mô hình chủ đề liên quan khác.

- Đề xuất mô hình HMSĐ BiLSTM-KD-NER chắt lọc tri thức mô hình học sâu cho bài toán nhận dạng thực thể y sinh tiếng Việt, xây dựng tập dữ liệu thực nghiệm và tiến hành thực nghiệm đánh giá mô hình. Kết quả thực nghiệm mô hình trên các kịch bản khác nhau cho thấy tính hiệu quả của mô hình HMSĐ BiLSTM-KD-NER.

Luận án công bố bốn bài báo trên các ấn phẩm Scopus và đã nhận được hai tham chiếu Scopus từ các tác giả nước ngoài.

Luận án vẫn còn các hạn chế sau đây:

- Mô hình CĐSĐ miền gần CD-AMC và các biến thể của nó dựa trên các chắt lọc tri thức được nhận biết khá dễ dàng. Các độ đo trong các định nghĩa miền gần mới chỉ dựa trên các độ đo tương tự đơn giản (dạng độ đo cosin), chưa đề cập tới các độ đo tương tự tinh vi hơn. Đồng thời, luận án chưa đề cập tới một số giải pháp cải tiến từ gợi ý trong nghiên cứu của Z. Chen (Z. Chen và cộng sự, 2015) như mẩu tri thức (khai phá tri thức must-link với nhiều hơn hai từ (tập mục phổ biến có từ ba mục/từ trở lên) hoặc khai phá mẫu hiếm cho chủ đề hẹp, sử dụng phân cụm các chủ đề tìm các chủ đề tương tự, v.v.).
- Chắt lọc tri thức mô hình trong BiLSTM-KD-NER chưa đạt được mức độ tinh vi như được mô tả trong (Isele và cộng sự, 2017 và Rostami và cộng sự, 2020].

Hướng phát triển tiếp theo của luận án là tập trung giải quyết một số hạn chế đã được chỉ ra. Nghiên cứu sinh sẽ tiến hành các công việc sau đây:

- Khai thác công cụ của Y. Papanikolaou và cộng sự (2017) ước tính tham số lấy mẫu Gibbs cải tiến mô hình LDA để xây dựng mô hình chủ đề A_{N+1} cho tập dữ liệu miền hiện tại D_{N+1} có ít dữ liệu. Phân tích sâu mô hình chủ đề nơ-ron suốt đời LNTM để cải thiện độ đo miền gần, đồng thời, xem xét việc cài đặt DocNADE trong việc cải thiện các chủ đề trong mô hình chủ đề cho bài toán hiện tại.
- Cải tiến khai phá must-link có ba từ trở lên hoặc must-link cho các chủ đề miền hẹp.
- Khảo sát bổ sung và sâu sắc hơn các nghiên cứu về chắt lọc tri thức cho học sâu suốt đời để đề xuất các kiểu chắt lọc tri thức mới.

Danh mục công trình khoa học của tác giả liên quan tới luận án

[NTCham1] Quang-Thuy Ha, Thi-Ngan Pham, Van-Quang Nguyen, Thi-Cham Nguyen, Thi-Hong Vuong, Minh-Tuoi Tran, Tri-Thanh Nguyen (2018). *A New Lifelong Topic Modeling Method and Its Application to Vietnamese Text Multi-label Classification*. ACIIDS 2018: 200-210 (Scopus, DBLP).

[NTCham2] Thi-Cham Nguyen, Thi-Ngan Pham, Minh-Chau Nguyen, Tri-Thanh Nguyen, Quang-Thuy Ha (2020). *A Lifelong Sentiment Classification Framework Based on a Close Domain Lifelong Topic Modeling Method*. ACIIDS 2020: 575-585 (Scopus, DBLP).

[NTCham3] Thi-Cham Nguyen, Thi-Ngan Pham, Hoang-Quynh Le, Tri-Thanh Nguyen, Hong-Nhung Bui, Quang-Thuy Ha (2020). *A Targeted Topic Model based Multi-Label Deep Learning Classification Framework for Aspect-based Opinion Mining*. IEEExplore KSE 2020: 165-170 (Scopus, DBLP).

[NTCham4] Thi-Cham Nguyen, Hoang-Quynh Le, Duy-Cat Can, Quang-Thuy Ha (2021). *Models Distillation with Lifelong Deep Learning for Vietnamese Biomedical Named Entity Recognition*. KSE 2021: 1-6 (Scopus, DBLP).