ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC - KỸ THUẬT MÁY TÍNH

GIAI ĐOẠN 1 (GĐ1):
ĐỀ CƯƠNG LUẬN VĂN/ ĐỒ ÁN CHUYÊN NGÀNH/ ĐỒ ÁN MÔN HỌC KỸ THUẬT MÁY TÍNH

Tên đề tài

Tiếng Việt: Rút trích quan hệ giữa Ý định và Thực thể sử dụng hướng tiếp cận Đọc hiểu Máy cho nhiệm vụ xây dựng Đồ thị Tri thức trong lĩnh vực Giáo dục.

Tiếng Anh: Relation Extraction between Intent and Entity using Machine Reading Comprehension(MRC) approach for Knowledge Graph Construction in Education domain.

Giáo viên hướng dẫn: Bùi Công Tuấn

SV thực hiện: Lưu Quốc Bình 2033009

Tp. Hồ Chí Minh, Tháng 05 Năm 2024

Contents

1 Giới thiệu

Báo cáo này trình bày các phương pháp trích xuất quan hệ sử dụng mô hình Hiểu đọc máy tính (Machine Reading Comprehension - MRC).

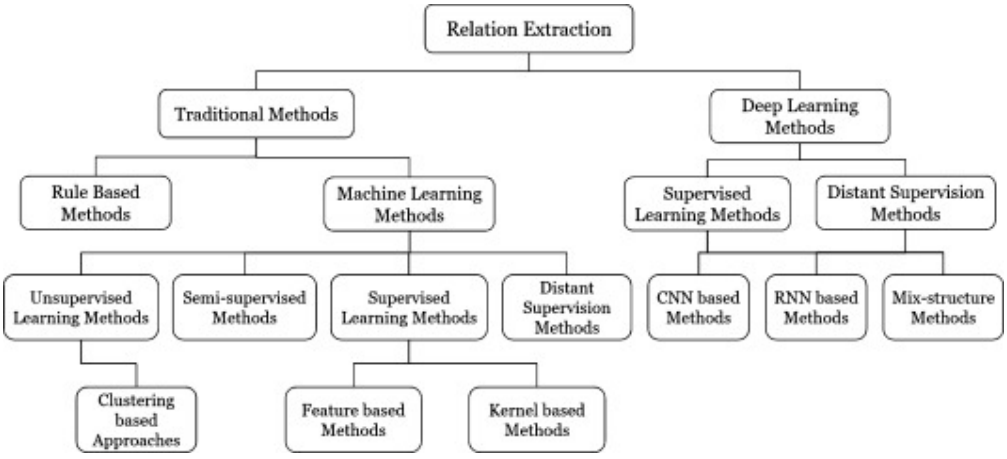
- Đường hướng nghiên cứu: Báo cáo tập trung vào phương pháp trích xuất quan hệ trong lĩnh vực giáo dục sử dụng MRC.

2 Phương pháp trích xuất quan hệ

2.1 Các phương pháp truyền thống

Như được thể hiện trong Hình 1, các phương pháp trích xuất quan hệ (RE) hiện có có thể được phân loại thành hai loại chính: Phương pháp truyền thống (Traditional Methods) và Phương pháp học sâu (Deep Learning Methods). Phương pháp truyền thống sử dụng các kỹ thuật dựa trên luật (Rule Based Methods) hoặc kỹ thuật học máy để trích xuất một tập hợp các quan hệ được xác định trước từ một kho văn bản (corpus). Các phương pháp truyền thống có thể được phân loại chi tiết thành phương pháp dựa trên luật (Rule Based Methods) và phương pháp học máy (Machine Learning Methods).

Các phương pháp học máy (Machine Learning Methods) có thể được phân loại thêm thành bốn loại: (1) Phương pháp không giám sát (Unsupervised Learning Methods). (2) Phương pháp có giám sát (Supervised Learning Methods). (3) Phương pháp bán giám sát (Semi-supervised). và (4) Phương pháp Giám sát từ xa (Distant supervision).



Nguồn: https://www.sciencedirect.com/science/article/pii/S2667305323000698br0060

Hình. 1: *Các loại phương pháp trích xuất quan hệ.*

- Phương pháp dựa trên luật (Rule-based): Sử dụng các quy tắc thủ công được thiết lập sẵn để trích xuất quan hệ.
- Phương pháp không giám sát (Unsupervised): Tự động học các mẫu trích xuất quan hệ từ dữ liệu.
- Phương pháp có giám sát (Supervised): Sử dụng dữ liệu được dán nhãn để huấn luyện mô hình trích xuất quan hệ.
- Phương pháp bán giám sát (Semi-supervised): Kết hợp dữ liệu được dán nhãn và không được dán nhãn để huấn luyện mô hình.
- Phương pháp giám sát từ xa (Distant supervision): Sử dụng dữ liệu gián tiếp để huấn luyện mô hình.

2.1.1 Phương pháp dựa trên luật (Rule-based)

Các phương pháp này còn được gọi là phương pháp mẫu thủ công (hand-built pattern methods). Chúng xác định một tập hợp các mẫu trích xuất (extraction patterns) cho các quan hệ được định nghĩa trước. Sau đó, các mẫu trích xuất này được đối chiếu với văn bản. Nếu một mẫu khớp với văn bản, thì một quan hệ tương ứng với mẫu đó được tìm thấy trong văn bản. Riloff (1993), Appelt et al. (1993), Agichtein và Gravano (2000), Jayram et al. (2006), Shen et al. (2007), Fundel et al. (2006), Nebhi (2013) đã sử dụng các phương pháp dựa trên luật để trích xuất quan hệ từ văn bản. Bảng 3 hiển thị các ví dụ về quy tắc để trích xuất quan hệ hạ đẳng (hyponymy) từ văn bản.

Pattern	Input sentence	Extracted relations
*such NP as {NP,}**	. . . works by such authors as Herrick, Goldsmith, and Shakespeare.	Hyponym (“author”, “Herrick”), Hyponym (“author”, “Goldsmith”), Hyponym (“author”, “Shakespeare”)
*NP {, NP} * {,} or other NP*	Bruises, wounds, broken bones or other injuries . . .	Hyponym (“bruise”, “injury”), Hyponym (“wound”, “injury”), Hyponym (“broken bones”, “injury”)

Hình. 2: *Ví dụ cho partern (mẫu/luật) Hyponyms (Hearst (1992)).*

Các phương pháp dựa trên luật đòi hỏi chuyên môn về lĩnh vực và kiến thức ngôn ngữ để xác định các mẫu trích xuất. Những phương pháp này phụ thuộc vào từng lĩnh vực cụ thể, trong đó cấu trúc của tài liệu là cố định và các quan hệ mục tiêu được xác định trước. Nếu chuyển từ lĩnh vực này sang lĩnh vực khác, thì cần phải xác định lại một tập hợp mới các quan hệ mục tiêu và mẫu trích xuất. Do đó, các phương pháp dựa trên luật đòi hỏi nhiều công sức thủ công và không thể sử dụng cho các tập văn bản đa dạng (heterogeneous corpora).

2.1.2 Phương pháp không giám sát (Unsupervised)

Phương pháp không giám sát không yêu cầu bất kỳ dữ liệu được dán nhãn nào. Hầu hết các phương pháp RE không giám sát sử dụng cách tiếp cận dựa trên cụm (clustering). Một trong những phương pháp RE không giám sát dựa trên cụm tiên phong được đề xuất bởi Hasegawa et al. (2004). Họ sử dụng công cụ đánh nhãn Thực thể Danh riêng (NE) để trích xuất các thực thể, cho phép tập trung chỉ vào các quan hệ với những thực thể được đề cập. Hình 2 minh họa các bước của phương pháp học không giám sát:

- (1) Nhận dạng các thực thể tên riêng trong kho văn bản.
- (2) Xác định các thực thể tên riêng đồng xuất hiện và ngữ cảnh của chúng.
- (3) Cụm các cặp thực thể dựa trên tính tương đồng ngữ cảnh.
- (4) Gán tên quan hệ ngữ nghĩa cho mỗi cụm.

Hình. 3: *Cách tiếp cận dựa trên phân cụm cho Unsupervised Learning.*

2.1.3 Phương pháp có giám sát (Supervised)

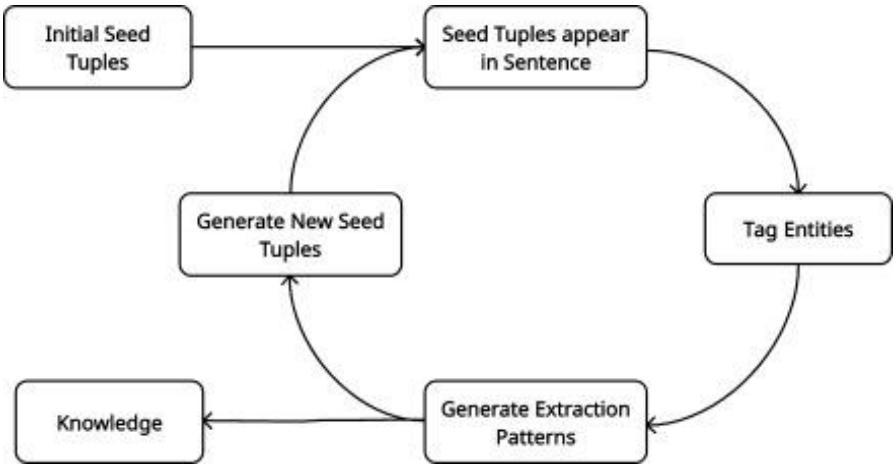
Phương pháp có giám sát yêu cầu một lượng lớn dữ liệu huấn luyện được dán nhãn với tập hợp các thực thể và quan hệ. Phương pháp này sử dụng dữ liệu huấn luyện để đào tạo bộ phân loại, trích xuất quan hệ từ dữ liệu kiểm tra. Có hai loại phương pháp có giám sát: Phương pháp dựa trên đặc trưng (Feature-based methods) và Phương pháp dựa trên Hạt nhân (Kernel-based methods) (Pawar et al. (2017)).

- Trong phương pháp dựa trên đặc trưng (Rink và Harabagiu (2010), Kambhatla (2004), Zhou et al. (2005)), một tập hợp các đặc trưng được tạo ra cho mỗi quan hệ trong dữ liệu huấn luyện, và bộ phân loại được đào tạo để trích xuất một thể hiện quan hệ mới. Một số đặc trưng về từ vựng (lexical), cú pháp (syntactic) và ngữ nghĩa (semantic) được mô tả trong (Kambhatla (2004)). Việc lựa chọn các đặc trưng ảnh hưởng đến hiệu suất của hệ thống RE có giám sát dựa trên đặc trưng.
- Không giống như phương pháp dựa trên đặc trưng, phương pháp dựa trên hàm nhân (Bunescu và Mooney, 2005b, Bunescu và Mooney, 2005a, Zelenko et al. (2003), Culotta và Sorensen (2004), Zhang et al. (2008), Zhang (2004)) không sử dụng các bước tiền xử lý để thiết kế đặc trưng. Trong phương pháp dựa trên hàm nhân, các kernel functions được sử dụng để xác định độ tương đồng giữa hai biểu diễn thể hiện quan hệ, trong khi Máy vectơ hỗ trợ (SVM) được sử dụng để phân loại. Hiệu suất của hệ thống RE dựa trên kernel functions phụ thuộc vào việc thiết kế các hàm nhân.

2.1.4 Phương pháp bán giám sát (Semi-supervised)

Việc tạo dữ liệu cho các phương pháp RE có giám sát đòi hỏi nhiều chi phí, nhân công và thời gian. Phương pháp bán giám sát tự động tạo dữ liệu được dán nhãn bằng thuật toán

bootstrapping. Cách tiếp cận này có hai ưu điểm: (1) giảm thiểu công sức cần thiết để tạo dữ liệu được dán nhãn; (2) tận dụng dữ liệu không được dán nhãn unlabeled data sẵn có miễn phí. Thuật toán bootstrapping yêu cầu một lượng lớn dữ liệu không được dán nhãn và một số ít instance hạt giống (seed instance) của kiểu quan hệ mong muốn. Ví dụ, để trích xuất quan hệ "Thủ đô của" (CapitalOf), các instance hạt giống "(New Delhi, Ấn Độ)", "(Canberra, Úc) và (London, Anh)" có thể được sử dụng để học một mẫu trích xuất. Với các instance hạt giống này làm đầu vào, thuật toán bootstrapping có thể trích xuất các quan hệ tương tự với các cặp thực thể như "(Paris, Pháp)". Hình 3 minh họa mô hình để trích xuất các mẫu và các cặp thực thể hạt giống bằng phương pháp học bán giám sát. KnowItAll (Etzioni et al. (2004)), TextRunner (Banko et al. (2007)), OLLIE (Mausam et al. (2012)), v.v. sử dụng phương pháp bán giám sát cho RE.



Hình. 4: *Phương pháp học bán giám sát (Semi-Supervised Learning) để trích xuất quan hệ.*

2.1.5 Phương pháp giám sát từ xa (Distant supervision)

Phương pháp này còn được gọi là phương pháp giám sát yếu (weakly supervised methods) hoặc phương pháp dựa trên kiến thức (knowledge based methods). Mintz et al. (2009) đã đề xuất phương pháp giám sát từ xa, trong đó dữ liệu huấn luyện được tạo tự động bằng cách liên kết văn bản với một Cơ sở tri thức (Knowledge Base - KB). Điều này giúp loại bỏ vấn đề dán nhãn thủ công cho dữ liệu huấn luyện. Học từ xa dựa trên giả định rằng nếu hai thực thể có quan hệ xuất hiện trong KB, thì tất cả các cụm từ đề cập đến hai thực thể này đều có thể diễn đạt mối quan hệ đó.

Theo cách này, Giám sát từ xa sử dụng cơ sở tri thức, chẳng hạn như Freebase, để trích xuất các mối quan hệ giữa hai đối tượng. Khi cùng một cặp thực thể xuất hiện trong cả một câu và một KB, thì theo quy tắc heuristic, câu đó được liên kết với mối quan hệ khớp từ KB. Ví dụ, hãy xem xét câu: ′Bill Gates là người sáng lập Microsoft.′ Nếu người ′Bill Gates′ và tổ chức ′Microsoft′ xuất hiện dưới dạng bộ ba ′(entity1: Bill Gates, entity2: Microsoft, relation: founder_of)′ trong Freebase, thì hai thực thể này đại diện cho mối quan hệ founder_of (người sáng lập). Smirnova và Cudré-Mauroux (2018) đã sử dụng một quy trình như được thể hiện trong Hình 4 để tạo dữ liệu huấn luyện bằng Giám sát từ xa. Các cách tiếp cận khác (Riedel et al. (2010), Takamatsu et al. (2012), Zeng et al. (2014), Qin et al. (2018), Chen and Manning (2014), Quirk and Poon (2017)) sử dụng Giám sát từ xa để trích xuất các mối quan hệ từ văn bản.

2.2 Phương pháp học sâu (Deep learning)

Mạng nơ-ron sâu (Deep Neural Networks - DNNs) (Bengio (2009)) đã trở nên phổ biến trong thập kỷ qua cho nhiều ứng dụng, bao gồm xử lý ngôn ngữ tự nhiên (NLP), thị giác máy tính (CV) và nhận dạng giọng nói. Kiến trúc DNN dựa trên Mạng nơ-ron tích chập (Convolutional Neural Network - CNN), Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN), Bộ nhớ dài hạn (Long Short-Term Memory - LSTM), Gated Recurring Unit (GRU), Bộ mã hóa song hướng từ Transformers (BERT) được sử dụng cho các nhiệm vụ NLP khác nhau chẳng hạn như đánh nhãn từ tính (POS tagging) (Gopalakrishnan et al. (2019), Srivastava et al. (2018), Kumar et al. (2018)), dịch máy (Wu et al. (2016), Wang et al. (2019), Gehring et al. (2017)), trả lời câu hỏi (Zhu et al. (2018), Lei et al. (2018), Liu et al. (2020)), gắn nhãn vai trò ngữ nghĩa (semantic role labelling) (He et al. (2017), Zhou and Xu (2015), Marcheggiani and Titov (2017)), sinh hội thoại (Li et al. (2016), Miranda and Kessaci (2020), Zhao and Eskenazi (2016)), sinh văn bản (Chen et al. (2020), Raffel et al. (2020), Li et al. (2021)), phân tích định hướng (sentiment analysis) (Yu et al. (2019), AL-Smadi et al. (2019), Zhao et al. (2018)), tóm tắt tự động (Zhang et al. (2019), See et al. (2017), Liu and Lapata (2019), Nallapati et al. (2016), Cai et al. (2019)) v.v.

2.2.1 CNN based methods

2.2.2 RNN and LSTM based methods

2.2.3 Encoder-decoder/transformer based methods

3 Các MRC phù hợp với Tiếng Việt

4 Định nghĩa các loại quan hệ được rút trích lĩnh vực giáo dục

5 Kết quả thử nghiệm

6 Danh sách công việc còn lại