


Họ và tên (IN HOA)	NGUYỄN QUAN DUY TÙNG
Ảnh	
Số buổi vắng	0
Bonus	11
Tên đề tài (VN)	XÂY DỰNG VÀ LÀM GIÀU CƠ SỞ TRI THỨC TỪ CÁC TẬP DỮ LIỆU TIẾNG VIỆT ĐƯỢC ĐỊNH DANH.
Giới thiệu	<ul style="list-style-type: none"> • <i>Bài toán/vấn đề</i> <p>Bài toán khai thác các quan hệ (r - relation) giữa các thực thể (E - entities) được trích xuất từ các bộ dữ liệu tiếng Việt được xây dựng theo định hướng định danh thực thể (NED - named entity disambiguation) để tạo và làm giàu cơ sở tri thức (KBs - knowledge bases) ở dạng đồ thị tri thức (KGs - knowledge graphs) có cấu trúc.</p> <ul style="list-style-type: none"> • <i>Lý do chọn đề tài, khả năng ứng dụng thực tế, tính thời sự</i>

Cơ sở tri thức (KB) ở dạng đồ thị tri thức (KG) là tài nguyên trong nhiều hệ thống hỏi đáp (Q&A), khuyến nghị, xử lý ngôn ngữ,... Nguồn tài nguyên dữ liệu dựa trên KB với ngôn ngữ là tiếng Việt có số lượng còn hạn chế và đa số định hướng để xây dựng các tập dữ liệu tiếng Việt tập trung vào việc thu thập thực thể và định danh với số lượng lớn. Tuy nhiên, quan hệ giữa các thực thể mới chính là tác nhân cung cấp ngữ cảnh để xác định đúng ngữ nghĩa của các thực thể đa nghĩa trong tiếng Việt. Ngoài ra việc kết hợp KB dựa trên bộ dữ liệu như phoNER_COVID19 trong các hệ thống tìm kiếm. KB là cơ sở để đánh giá thông tin sai, tin giả nhờ vào đối chiếu quan hệ giữa các thực thể được trích xuất và định danh trong tập dữ liệu.

- *Input và output*

Input: một câu $S = \langle w_1, w_2, \dots, w_i \rangle$ với w_i là từ ở vị trí i trong câu.

Output: trích xuất các bộ ba có cấu trúc $O = \langle o_1, o_2, \dots, o_j \rangle$ từ các câu, trong đó $o_j = \langle h_j, r_j, t_j \rangle$ với h_j, t_j thuộc tập thực thể (E) và r_j thuộc tập quan hệ (R).

- *Ví dụ minh họa*

Input: "Đại học công nghệ thông tin (UIT) là trường đại học công lập ở thành phố Hồ Chí Minh (tpHCM)."

Output: (phương pháp tiếp cận không giám sát)

$\langle \text{UIT, là, trường đại học công lập} \rangle$

$\langle \text{UIT, là trường đại học công lập ở, tpHCM} \rangle$

Output: (phương pháp tiếp cận có giám sát)

$\langle \text{UIT, thực thể của, trường đại học công lập} \rangle$

	<p><UIT, nằm ở, tpHCM ></p> <p>Output: (đầu ra được chuẩn hóa)</p> <p><Q12345, P12, Q23456></p> <p><Q12345, P123, Q54321></p>																																								
Mục tiêu	<ul style="list-style-type: none">• Áp dụng trích xuất quan hệ end-to-end với mô hình neural encoder - decoder, n-gram based, attention với các cụm từ trong câu.• Gán ID cho các thực thể, quan hệ tối ưu không gian lưu trữ. Làm giàu, tinh gọn tập quan hệ nhằm tăng tính chính xác của KB.• Xây dựng và làm giàu cơ sở tri thức (KB) có cấu trúc với ngôn ngữ tiếng Việt.																																								
Nội dung và phương pháp thực hiện	<div><div><p>Dataset Collection Module</p><p>Wikipedia article</p><p>Wikidata</p><p>Distant supervision</p></div><div><p>Embedding Module</p><p>Joint learning skip-gram & TransE</p><p>Word Embeddings</p><table><tr><td>0.2</td><td>0.4</td><td>0.1</td><td>0.2</td><td>0.1</td></tr><tr><td>0.1</td><td>0.5</td><td>0.1</td><td>0.1</td><td>0.2</td></tr><tr><td>0.2</td><td>0.3</td><td>0.3</td><td>0.3</td><td>0.2</td></tr><tr><td>0.4</td><td>0.2</td><td>0.2</td><td>0.1</td><td>0.1</td></tr></table><p>Entity Embeddings</p><table><tr><td>0.1</td><td>0.5</td><td>0.1</td><td>0.4</td><td>0.2</td></tr><tr><td>0.1</td><td>0.5</td><td>0.1</td><td>0.5</td><td>0.1</td></tr><tr><td>0.2</td><td>0.3</td><td>0.3</td><td>0.3</td><td>0.3</td></tr><tr><td>0.2</td><td>0.3</td><td>0.3</td><td>0.3</td><td>0.1</td></tr></table></div><div><p>Neural Relation Extraction Module</p><p>Expected output: <Q12345, P12, Q23456> <Q12345, P123, Q54321></p><p>Triple classifier</p><p>Decoder</p><p>N-gram-based attention</p><p>Encoder</p></div></div> <p>Sentence input: Đại học công nghệ thông tin là trường đại học công lập ở thành phố Hồ Chí Minh.</p> <div><p>Sentence-Triple pairs</p><p>Sentence input: Đại học công nghệ thông tin là trường đại học công lập ở thành phố Hồ Chí Minh.</p><p>Expected output: Q12345 P12 Q23456 P123 Q54321</p></div>	0.2	0.4	0.1	0.2	0.1	0.1	0.5	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.1	0.1	0.1	0.5	0.1	0.4	0.2	0.1	0.5	0.1	0.5	0.1	0.2	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.1
0.2	0.4	0.1	0.2	0.1																																					
0.1	0.5	0.1	0.1	0.2																																					
0.2	0.3	0.3	0.3	0.2																																					
0.4	0.2	0.2	0.1	0.1																																					
0.1	0.5	0.1	0.4	0.2																																					
0.1	0.5	0.1	0.5	0.1																																					
0.2	0.3	0.3	0.3	0.3																																					
0.2	0.3	0.3	0.3	0.1																																					

- Module thu thập dữ liệu

- Mục đích là trích xuất các bộ ba từ các câu với mô hình học có giám sát trên bộ dữ liệu được gán nhãn như *Wikidata* (<https://dumps.wikimedia.org/viwiki/latest/viwiki-latest-pages-articles.xml.bz2>), *VLSP 2016 NER*, *VLSP 2018 NER*, . . . bằng cách ánh xạ thực thể, quan hệ được tìm thấy. Kết hợp mô hình distant supervision gán ID cho chúng.
- Trích xuất các câu chứa implicit entity được nắm bắt bởi co-reference resolution trong tập dữ liệu làm giàu được định danh.
- Paraphrase detection lập từ điển chứa các vị từ quan hệ trích xuất từ các tập được định danh. Từ điển vị từ dùng để so khớp các quan hệ, đồng thời giúp loại bỏ những câu không chứa bất cứ quan hệ.

- Embedding module

- Nhúng đồng thời từ và thực thể được định danh để nắm bắt sự giống nhau giữa chúng.
- Phép nhúng dựa vào việc tối tiểu hàm mục tiêu margin-based

$$J_E = \sum_{t_r \in T_r} \sum_{t'_r \in T'_r} \max(0, [\gamma + f(t_r) - f(t'_r)]) \quad (1)$$

$$T_r = \{\langle h, r, t \rangle | \langle h, r, t \rangle \in G\} \quad (2)$$

$$T'_r = \{\langle h', r, t \rangle | h' \in E\} \cup \{\langle h, r, t' \rangle | t' \in E\} \quad (3)$$

$$f(t_r) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \quad (4)$$

với $\|\mathbf{x}\|$ là chuẩn L1 của vector \mathbf{x} , γ là tham số margin, T_r là tập bộ ba trong KB, T'_r là tập các bộ ba đứt gãy tìm thấy trong câu.

- Quá trình trên giúp tạo lập kho văn bản bằng cách kết hợp văn bản gốc và câu mà các thực thể được gán ID. Ví dụ như câu "*UIT là trường đại học công lập ở tpHCM*" được chuyển thành "*Q12345 là Q23456 ở Q54321*".
- Sau đó, sử dụng *skip-gram* và *TransE* để trích xuất ngữ cảnh bằng cách nhúng câu $[w_1, w_2, \dots, w_n]$ tối thiểu hàm mục tiêu J_W

$$J_W = \frac{1}{T} \sum_{t=1}^n \sum_{-c \leq j \leq c, j \neq 0} \log P(w_{t+j}|w_t) \quad (5)$$

$$P(w_{t+j}|w_t) = \frac{\exp(\mathbf{v}'_{w_{t+j}} \mathbf{v}_{w_t})}{\sum_{i=1}^W (\mathbf{v}'_i \mathbf{v}_{w_t})} \quad (6)$$

- Hàm mục tiêu tổng thể $J = J_W + J_E$ (7)
- Ví dụ câu "*Q12345 là Q23456 ở Q54321*" trở thành "*Q12345 P12 Q23456 P123 Q54321*".
- Neural relation extraction (RE) module
 - Sử dụng mô hình encoder-decoder, LSTM networks chuyển các câu thành chuỗi các bộ ba.
 - Encoder-decoder với attention model không thể nắm bắt quan hệ giữa các từ có cùng định danh nhưng khác đối tượng, ví dụ như "Đại học CNTT" và "Khoa CNTT". Điều này có thể gây ra lỗi trong quá trình gán ID.
 - Thay vào đó sử dụng mô hình Attention dựa trên N-gram, bằng cách tính toán attention weights của các cặp n-gram trong câu đầu vào với vector ngữ cảnh ($N = 3$)

	$\mathbf{c}_t^d = \left[\mathbf{h}^e; \sum_{n=1}^{ N } \mathbf{W}^n \left(\sum_{i=1}^{ X^n } \alpha_i^n \mathbf{x}_i^n \right) \right] \quad (8)$ $\alpha_i^n = \frac{\exp(\mathbf{h}^{e\top} \mathbf{V}^n \mathbf{x}_i^n)}{\sum_{j=1}^{ X^n } \exp(\mathbf{h}^{e\top} \mathbf{V}^n \mathbf{x}_j^n)} \quad (9)$ <p>với \mathbf{c} là vector ngữ cảnh, \mathbf{x} là vector từ nhúng, \mathbf{W} và \mathbf{V} là ma trận tham số, α là attention weights.</p>
Kết quả dự kiến	<ul style="list-style-type: none"> • So sánh giữa các phương pháp <i>CNN (SOTA với phương pháp học có giám sát), MiniE (SOTA với phương pháp học không giám sát), ClauseIE, với độ đo precision, recall, F1.</i> • Bộ dữ liệu <ul style="list-style-type: none"> ◦ <i>VLSP 2016 NER, VLSP 2018 NER, PhoNER_COVID19</i>
Tài liệu tham khảo	<p>[1] Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. Neural relation extraction for knowledge base enrichment. In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i>, pages 229–240, 2019.</p> <p>[2] Huyen TM Nguyen, Quyen T Ngo, Luong X Vu, Vu M Tran, and Hien TTNguyen. <i>Vlsp shared task: Named entity recognition. Journal of Computer Science and Cybernetics</i>, 34(4): 283–294, 2018.</p> <p>[3] Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. <i>More data, more relations, more context and more openness: A review and outlook for relation extraction. arXiv preprint arXiv: 2004.03186</i>, 2020.</p>