


Họ và tên (IN HOA)	TRẦN VĂN SAN MSHV: 2001013
Ảnh	
Số buổi vắng	0
Bonus	22
Tên đề tài (VN)	HỌC KHÔNG GIAN CON TỪ VỰNG TRONG KHÔNG GIAN VECTOR PHÂN PHỐI
Tên đề tài (EN)	LEARNING LEXICAL SUBSPACES IN A DISTRIBUTIONAL VECTOR SPACE
Giới thiệu	<ul style="list-style-type: none"> • <i>Bài toán/vấn đề mà đề tài muốn giải quyết</i> <p>Xây dựng một framework học các không gian con tuyến tính từ vừng trong không gian vector phân phối. Framework này có thể mô hình hóa tất cả các loại quan hệ từ vựng-ngữ nghĩa cốt lõi, cụ thể là: hút đối xứng (symmetric attract) như từ đồng nghĩa (synonymy) và đẩy đối xứng (symmetric repel) như từ trái nghĩa (antonymy), quan</p>

hệ bất đối xứng (asymmetric) như ẩn dụ (hypernymy) và hoán dụ (meronymy).

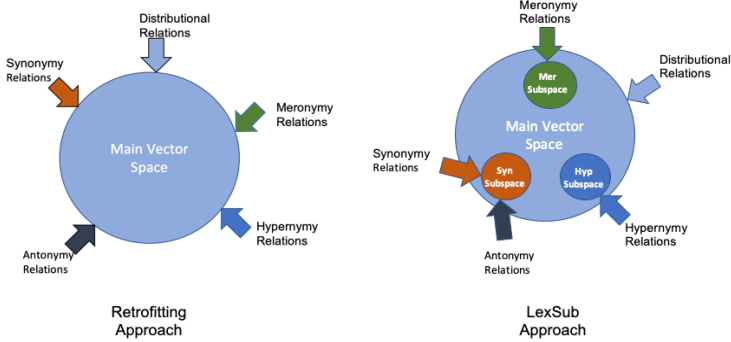
- *Lí do chọn đề tài, khả năng ứng dụng thực tế, tính thời sự*

- Nhúng từ (word embedding) bước tiền huấn luyện là nền tảng của kiến trúc xử lý ngôn ngữ tự nhiên hiện đại. Sự thành công của các phép nhúng từ bước tiền huấn luyện được cho là do khả năng đưa vào các giả thiết phân phối của chúng. Một số kỹ thuật đã được đề xuất trong các nghiên cứu nhằm sửa đổi các vector từ để kết hợp các quan hệ từ vựng - ngữ nghĩa vào không gian nhúng như Faruqui [1]. Cách tiếp cận của công trình này là đi sửa đổi không gian vector phân phối ban đầu. Hạn chế trong cách tiếp cận này là việc thay đổi không gian phân phối ban đầu có thể làm mất thông tin phân phối có được từ các vector hữu ích ban đầu, dẫn đến giảm hiệu suất khi được sử dụng trong các tác vụ xử lý ngôn ngữ tự nhiên.
- Một số cách tiếp cận nhằm mô hình hóa các loại quan hệ từ vựng khác nhau bao gồm: hút đối xứng, đẩy đối xứng, bất đối xứng. Các nghiên cứu được đề cập chỉ có thể mô hình hóa được một số kiểu quan hệ từ vựng. Ví dụ, Tifrea [2] chỉ có thể mô hình cho hoán dụ trong khi Vulic [3] có thể mô hình hoá quan hệ từ đồng nghĩa, từ trái nghĩa và ẩn dụ.
- Ta đề xuất một cách tiếp cận mới, được gọi là LEXSUB [4], với cách tiếp cận này nhằm hợp nhất ngữ nghĩa từ vựng và ngữ nghĩa phân phối. Phương pháp này có thể mô hình hóa tất cả các loại quan hệ từ vựng, cụ thể là, hút đối xứng, đẩy đối xứng và bất đối xứng, và sử dụng tất cả bốn quan hệ từ vựng chính được tìm thấy trong bộ dữ liệu WordNet của Miller [5] là: đồng nghĩa, trái nghĩa, ẩn dụ và hoán dụ.

- *Mô tả input và output, hình minh họa*

Input: Cho một tập gồm n từ vựng $V = \{x_1, x_2, x_3, \dots, x_n\}$.

Output: Các không gian con của quan hệ từ loại: Syn Subspace của quan hệ đồng nghĩa (Synonymy Relations) và quan hệ trái nghĩa (Antonymy Relations), Hyp Subspace của quan hệ ẩn dụ (Hypernymy Relations) và Mer Subspace của quan hệ hoán dụ (Meronymy Relations).

	 <p style="text-align: center;">Retrofitting Approach LexSub Approach</p>
Mục tiêu	<ul style="list-style-type: none"> - Mục tiêu 1: Xây dựng tri thức về các quan hệ từ vựng-ngữ nghĩa vào các phép nhúng từ phân phối bằng cách xác định các không gian con của không gian vector phân phối (distributional vector space) mà trong đó một quan hệ từ vựng cần giữ. - Mục tiêu 2: Đề xuất framework có thể mô hình hóa tất cả các loại quan hệ từ vựng, cụ thể là bốn quan hệ từ vựng cơ bản được phát hiện trên bộ từ vựng WordNet gồm: từ đồng nghĩa, trái nghĩa, ẩn dụ và hoán dụ. - Mục tiêu 3: Trong một bộ các tiêu chuẩn bên trong với cách tiếp cận này vượt trội hơn các phương pháp tiếp cận trước đây về các tác vụ quan hệ, phân loại và phát hiện từ ẩn dụ, đồng thời cũng hoàn thiện hơn các tác vụ từ tương tự. Không những thế nó cũng hoạt động tốt hơn các hệ thống trước đó về các tác vụ phân lớp bên ngoài từ việc sử dụng các tính chất quan hệ từ vựng.
Nội dung và phương pháp thực hiện	<p>Để đạt mục tiêu 1:</p> <ul style="list-style-type: none"> • Nội dung: Cho một tập gồm n từ vựng $V = \{x_1, x_2, x_3, \dots, x_n\}$ mục tiêu tạo ra một tập các vector $\{y_1, y_2, y_3, \dots, y_n\} \in \mathbb{R}^d$ tương ứng giống nhau cả về phân phối cũng như các quan hệ ngữ nghĩa - từ vựng. Ta gọi các vector này là các phép nhúng không gian vector chính. Xây dựng các quan hệ R trên không gian vector tương ứng với quan hệ từ vựng - ngữ nghĩa r. Các phần tử của tập quan hệ này là các cặp từ có thứ tự $(x_i, x_j) \in V \times V$; nghĩa là, nếu $(x_i, x_j) \in R$, thì x_i và x_j liên quan với nhau bằng quan hệ từ vựng r. Đối với các quan hệ đối xứng như đồng nghĩa và trái nghĩa, $(x_i, x_j) \in R$ thì $(x_j, x_i) \in R$. Tương tự, đối với các quan hệ

	<p>bất đối xứng như ẩn dụ và hoán dụ, x_j liên quan đến x_i bằng quan hệ r nếu $(x_i, x_j) \in R$ và $(x_j, x_i) \notin R$.</p> <ul style="list-style-type: none"> • Phương pháp: Phương pháp nghiên cứu định lượng. <p>Để đạt mục tiêu 2:</p> <ul style="list-style-type: none"> • Nội dung: Cấu trúc của framework gồm hai thành phần, nội dung thực hiện cho từng thành phần như sau: <ul style="list-style-type: none"> - Thành phần đầu tiên giúp mô hình học các không gian con từ vựng ở trong không gian vector phân phối. Các không gian con này được học bằng cách sử dụng hàm loss L_{lex} được định nghĩa bằng tổng các hàm loss trên các không gian con. - Thành phần thứ hai giúp mô hình học không gian vector phân phối. Việc huấn luyện không gian vector này được học bởi hàm loss L_{dist} được định nghĩa bằng việc bổ sung vào tập vector nhằm giảm thiểu thay đổi khoảng cách trong L2 giữa các từ nhúng. Do đó, tổng hàm loss tối ưu được định nghĩa là: $L_{total} = L_{dist} + L_{lex}$. • Phương pháp: Phương pháp nghiên cứu định lượng. <p>Để đạt mục tiêu 3:</p> <ul style="list-style-type: none"> • Nội dung: Thực hiện tác vụ bên trong và tác vụ bên ngoài trên các bộ dữ liệu đánh giá kết quả thực nghiệm nhằm so sánh với các cách tiếp cận trước đây: <ul style="list-style-type: none"> - Tác vụ bên trong (Intrinsic Task): Sử dụng bộ dữ liệu men3k của Bruni [6] để kiểm tra độ giống từ phổ biến để đánh giá độ tương tự của từ. Sử dụng bộ dữ liệu WordSim353 của Agirre [7] để đo khả năng lưu giữ thông tin phân phối của nhúng. Để đánh giá từ ẩn dụ được phân loại, tác giả sử dụng bộ dữ liệu Hyperlex của Gerz [8]. - Tác vụ bên ngoài (Extrinsic Tasks): Sử dụng bộ dữ liệu trả lời câu hỏi SQuAD1.1 của Rajpurkar [9] và sử dụng mô hình BiDAF của Seo [10] cho tác vụ trả lời câu hỏi. - Phương pháp: Đối với các mô hình trên, chúng ta sử dụng bằng cách triển khai tham chiếu trên các mô hình trong bộ
--	--

	<p>công cụ AllenNLP của Gardner [11]. Ta thay thế lớp đầu vào của các mô hình này bằng các nhúng mà ta muốn đánh giá. Chúng ta sẽ sử dụng hai thiết lập khác nhau cho các thí nghiệm bên ngoài của mình và báo cáo kết quả cho cả hai.</p> <ul style="list-style-type: none"> • Phương pháp: Phương pháp nghiên cứu thực nghiệm.
Kết quả dự kiến	<p><i>So sánh giữa các phương pháp</i></p> <p>Kết quả so sánh giữa LEXSUB và các phép nhúng trước đây được huấn luyện trên cùng bộ dữ liệu nguồn từ vựng. Cách tiếp cận LEXSUB hoạt động tốt hơn hoặc đầy đủ hơn so với các cách tiếp cận trước đây trên cùng bộ dữ liệu về đánh giá trên các tác vụ bên trong và sự vượt trội của nó trên một loạt các tác vụ bên ngoài có thể được sử dụng từ việc khai thác thông tin quan hệ từ vựng.</p> <ul style="list-style-type: none"> • <i>Bộ dữ liệu được sử dụng</i> <ul style="list-style-type: none"> - Bộ dữ liệu huấn luyện: GloVe [12] với 300 chiều được huấn luyện trên 6 tỷ mã token, bộ dữ liệu Wikipedia 2014. Kích thước từ vựng cho các nhúng GloVe là 400.000. - Bộ dữ liệu nguồn: Sử dụng WordNet làm cơ sở dữ liệu từ vựng cho tất cả các thử nghiệm. Chúng ta xem xét tất cả bốn loại quan hệ từ vựng: từ đồng nghĩa, trái nghĩa, từ ẩn dụ và từ hoán dụ. Chỉ những quan hệ bộ ba mà cả hai từ xuất hiện trong từ vựng mới được xem xét. Ta coi cả các từ ẩn dụ và khái niệm cho các quan hệ ẩn dụ, và cho các quan hệ hoán dụ, bộ phận, nội dung, cũng như các từ hoán dụ thành phần phụ dưới dạng các ràng buộc.
Tài liệu tham khảo	<p>[1] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In <i>Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i>, pages 1606–1615.</p>

- | | |
|--|--|
| | <p>[2] Alexandru Tifrea, Gary Becigneul, and Octavian- Eugen Ganea. 2018. Poincare GloVe: Hyper-bolic Word Embeddings. <i>arXiv:1810.06546 [cs]</i>.</p> <p>[3] Ivan Vulic and Nikola Mrksic. 2017. Specialising Word Vectors for Lexical Entailment. <i>arXiv: 1710.06371 [cs]</i>.</p> <p>[4] Kushal Arora, Aishik Chakraborty, Jackie C. K. Cheung. 2020. Learning Lexical Subspaces in a Distributional Vector Space. <i>Transactions of the Association for Computational Linguistics</i>, vol. 8, pp. 311–329, 2020.</p> <p>[5] George A. Miller. 1995. WordNet: A Lexical Database for English. <i>Communications of ACM</i>, 38(11):39–41.</p> <p>[6] Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. <i>Journal of Artificial Intelligence Research</i>, 49(1):1–47.</p> <p>[7] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In <i>Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09</i>, page 19, Boulder, Colorado. Association for Computational Linguistics.</p> <p>[8] Daniela Gerz, Ivan Vulic, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> pages 2173–2182, Austin, Texas. Association for Computational Linguistics.</p> <p>[9] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i>, pages 2383–2392.</p> |
|--|--|

	<p>[10] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. <i>arXiv:1611.01603 [cs]</i>.</p> <p>[11] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In <i>Proceedings of the 2018 Conference of the North American</i>.</p> <p>[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In <i>Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i>, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.</p>
--	---