BÀI 10 – BIỂU DIỄN TỪ

AI Academy Vietnam

Thông tin giảng viên

- TS. Phan Việt Anh
- Email: anhpv@lqdtu.edu.vn
- Mobile: 0975 639 757

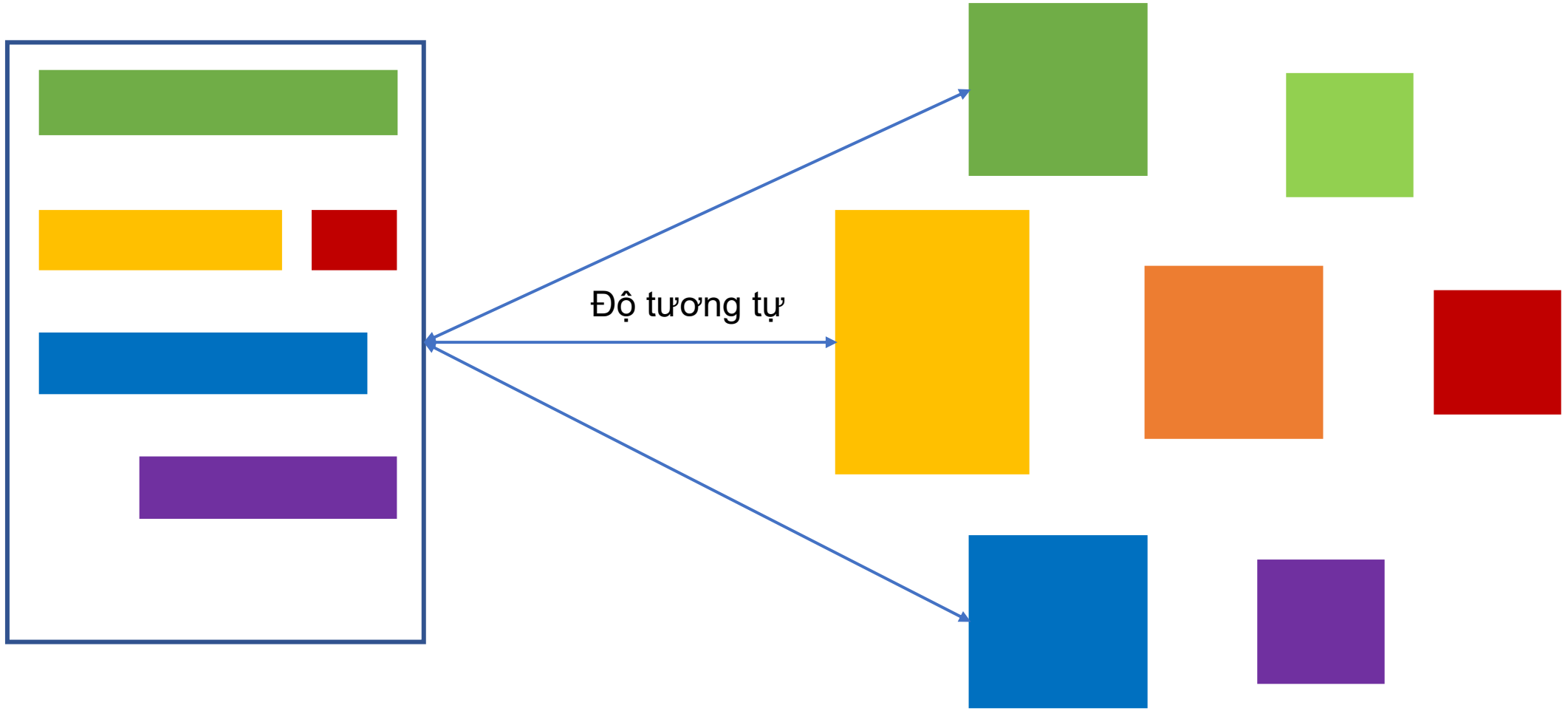
- Research interest: NLP, machine learning, deep learning.

Nội dung buổi học

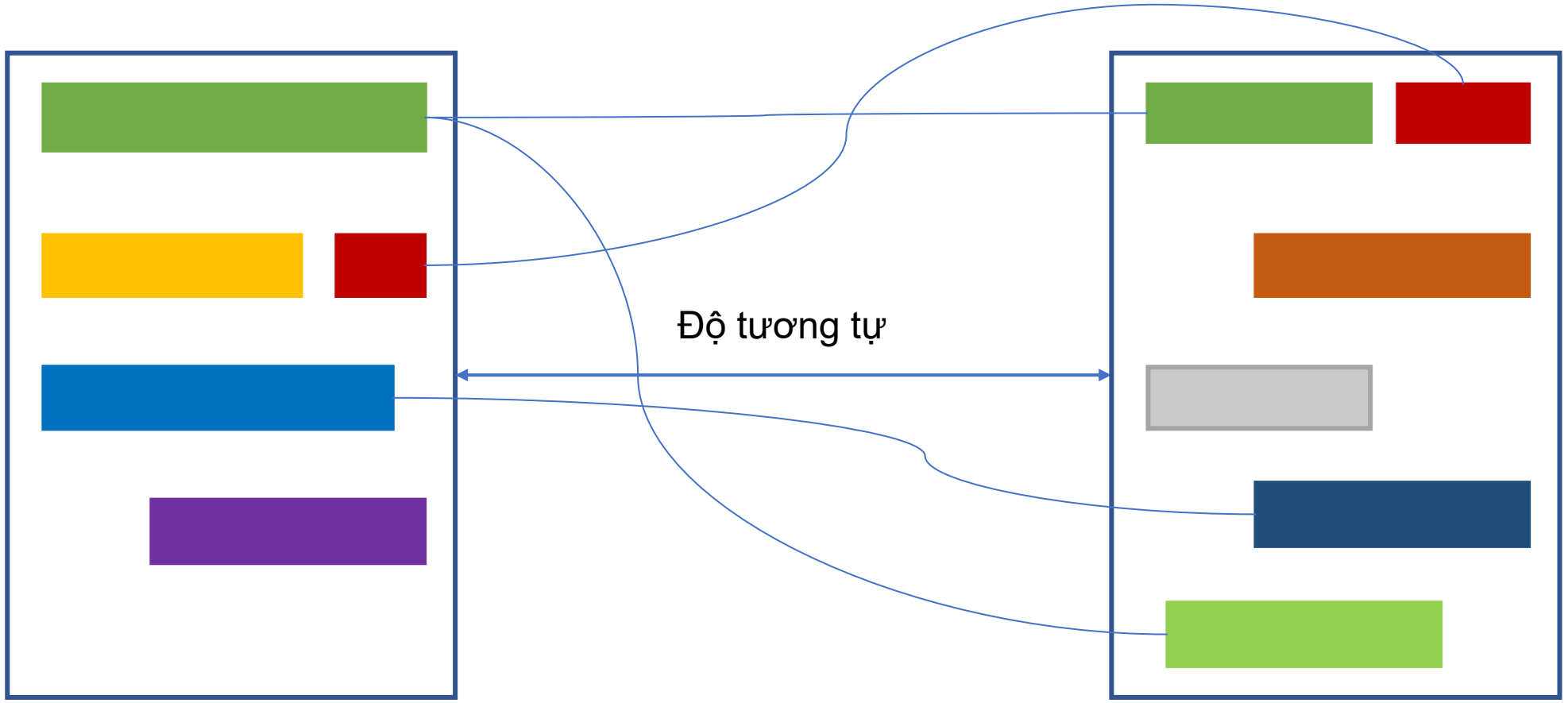
- Giới thiệu bài toán phát hiện đạo văn
- Nhắc lại một số kiến thức về vector hoá và word2vec
- Một số phương pháp biểu diễn khác
  - Glove
  - FastText
  - ELMo
- Xây dựng mô hình biểu diễn từ
- Ứng dụng biểu diễn từ cho bài toán phát hiện đạo văn

BÀI TOÁN PHÁT HIỆN ĐẠO VĂN

Bài toán phát hiện đạo văn



So sánh sự tương tự của 2 văn bản



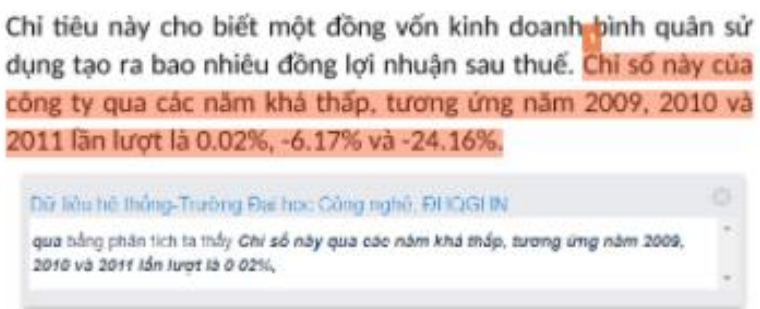
DoIT

TRANG CHỦ | TRỢ GIÚP | LIÊN HỆ | ĐĂNG NHẬP | ĐĂNG KÝ

Hệ thống hỗ trợ nâng cao chất lượng tài liệu

- ☑ Kiểm tra trùng lặp
- ☑ Sửa lỗi chính tả tiếng Việt

Dùng thử



Kiểm tra trùng lặp văn bản

Dữ liệu của hệ thống bao gồm các luận văn, khoá luận của nhiều trường đại học lớn trong cả nước, cũng như các bài báo, tạp chí và nhiều nguồn tài liệu uy tín khác.

Thách thức

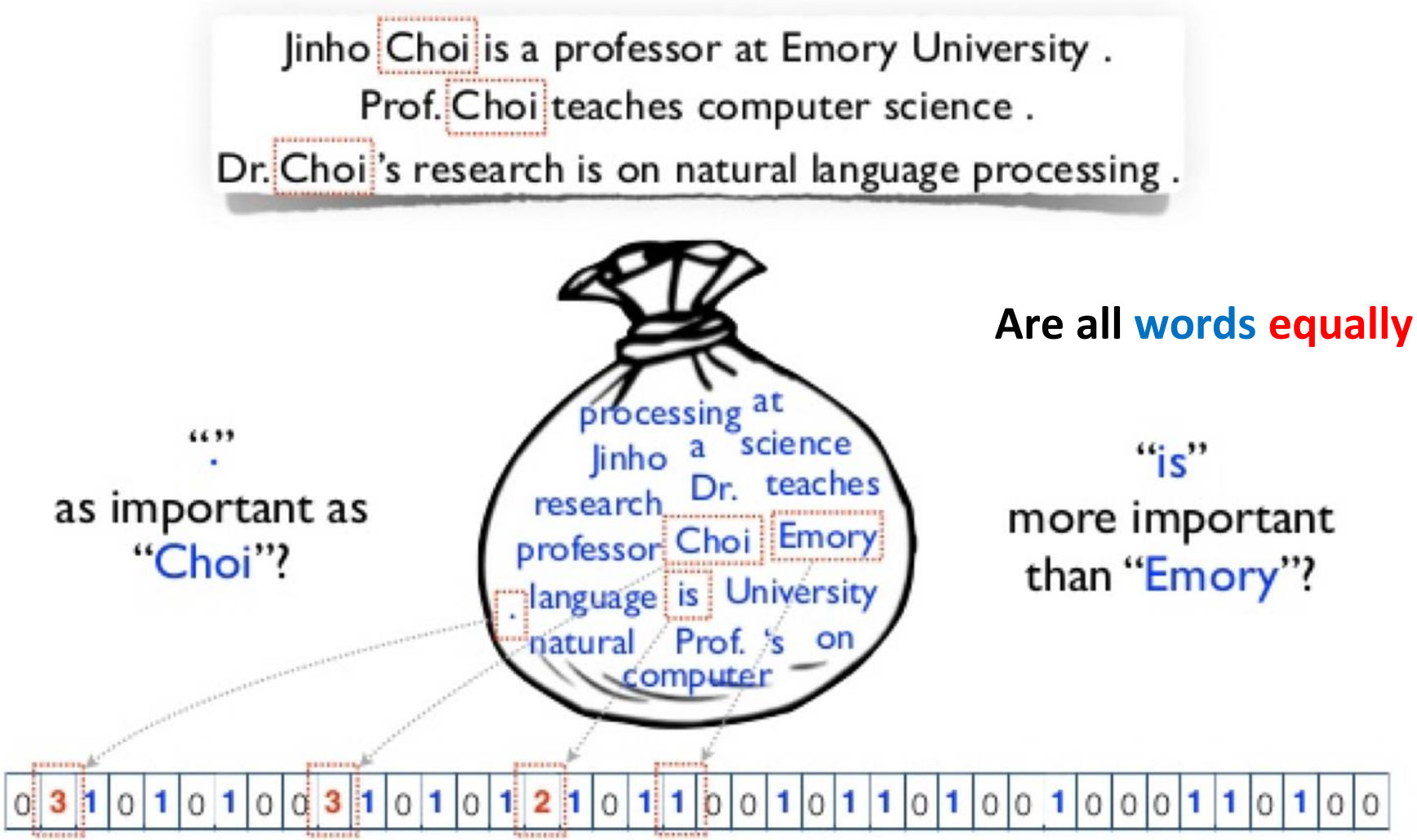
- Là một bài toán khó của thu hồi thông tin (information retrieval)
  - Bài toán nhỏ hơn là độ tương đồng trên mức câu
- Văn bản dài
- Sự đa dạng về cấu trúc và từ vựng
- Cần quan tâm tới
  - Ngữ nghĩa (semantics)
  - Cú pháp (syntactics)

NHẮC LẠI MỘT SỐ KIẾN THỨC

Một số phương pháp biểu diễn

- Mô hình túi từ (BoW)
- Sử dụng ma trận đồng xuất hiện
- TF-IDF
- Vector một giá trị (one-hot encoding)
- Word2Vec

Mô hình túi từ (bag-of-words model)



Are all words equally important?

Ma trận đồng xuất hiện

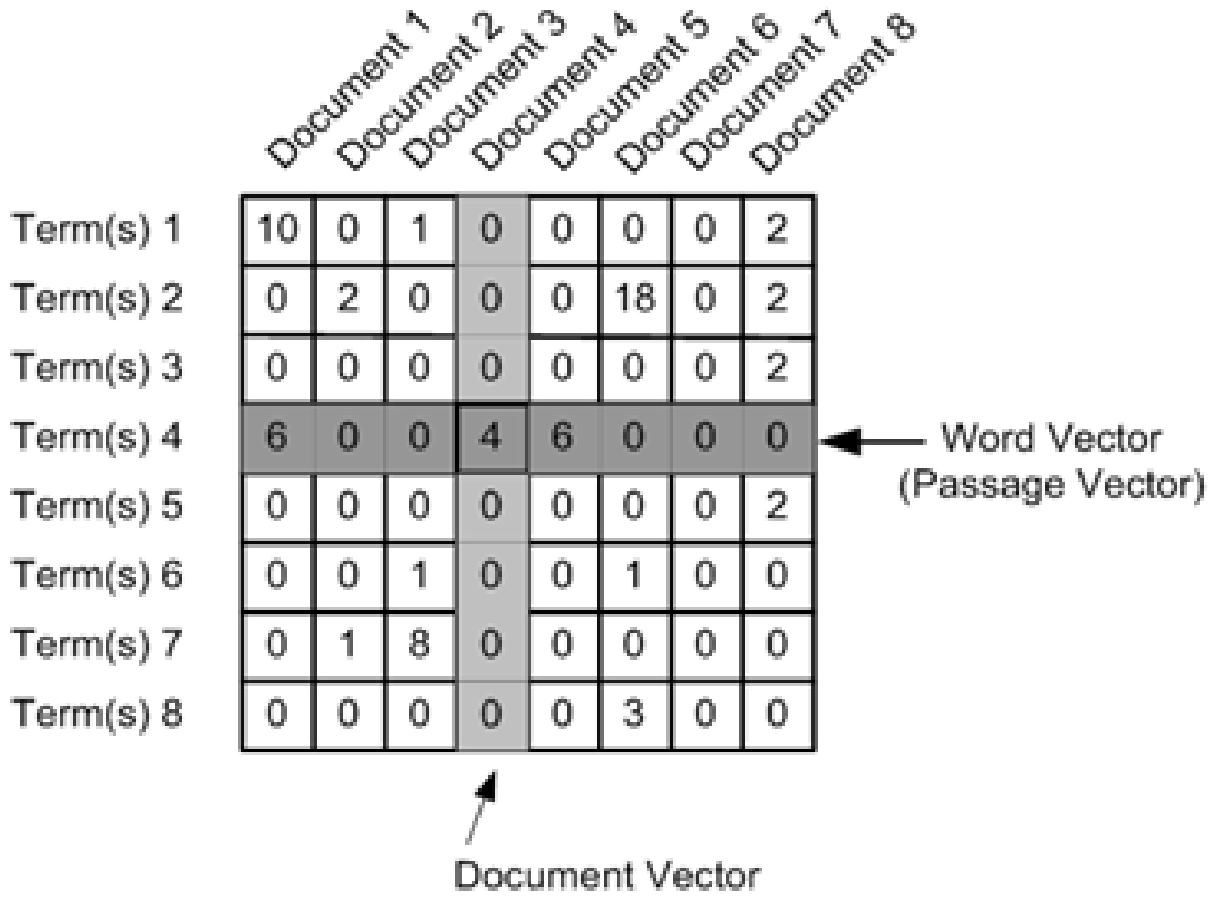
- Sử dụng một cửa sổ trượt
- Giá trị tại 1 ô
  - Sự đồng xuất hiện của 1 từ ở một hàng với các từ ở các cột

$$X = \begin{array}{c} \\ I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{array} \begin{array}{c} \begin{array}{cccccccc} I & like & enjoy & deep & learning & NLP & flying & . \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

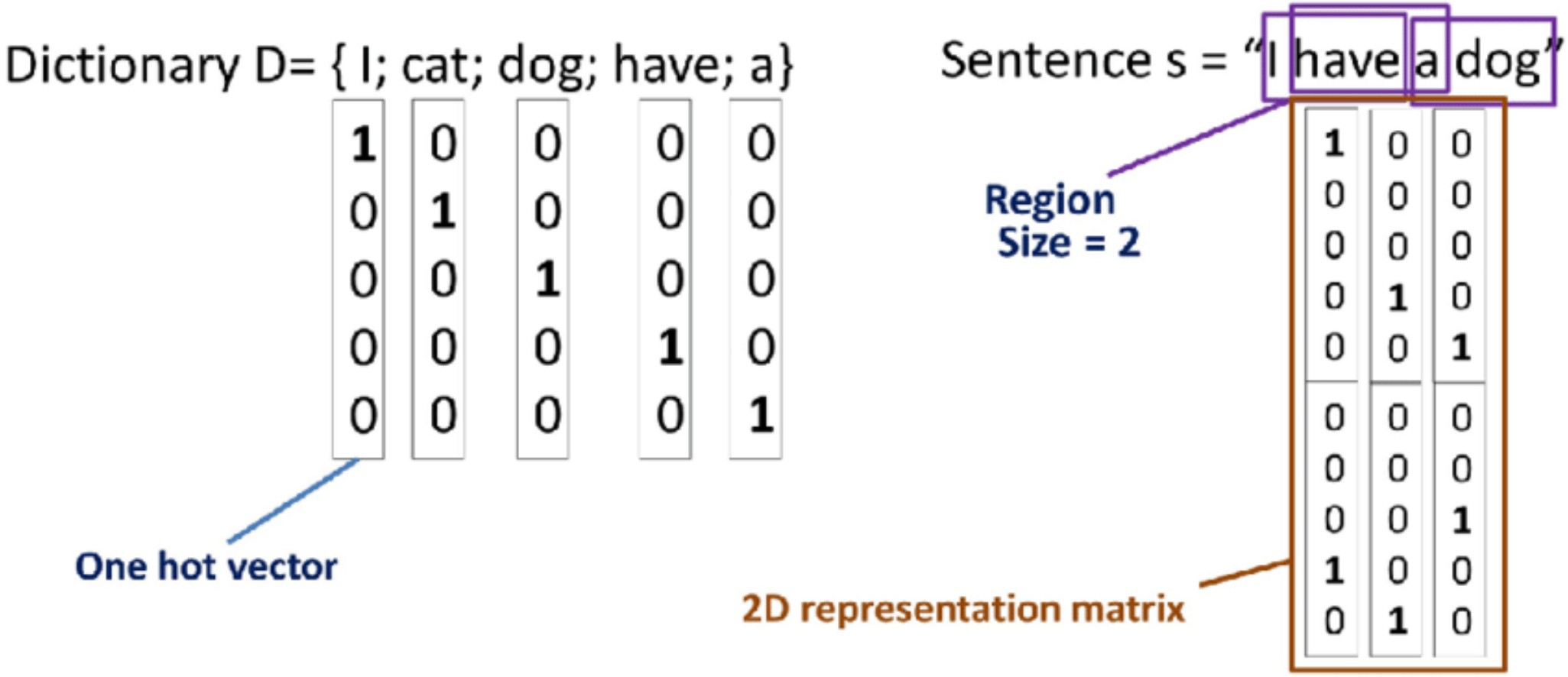
Mô hình TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

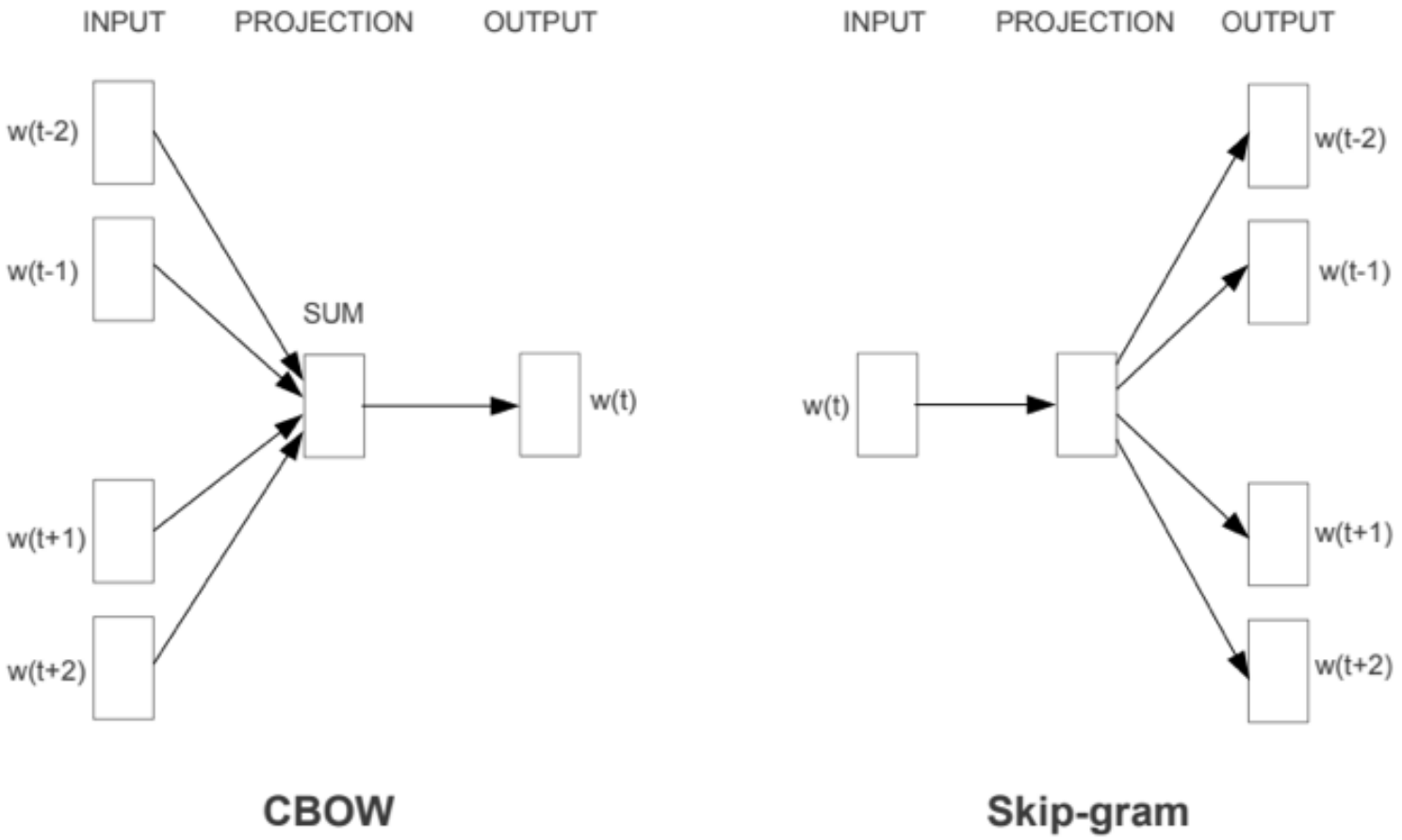


Vector 1 giá trị (one-hot vector)



Học biểu diễn Word2Vec: CBOW vs. Skip-gram

"Vec tơ từ mã hóa mối quan hệ ngữ nghĩa giữa các từ"



Efficient Estimation of Word Representations in Vector Space (Tomas et al., 2013)

Trực quan hoá

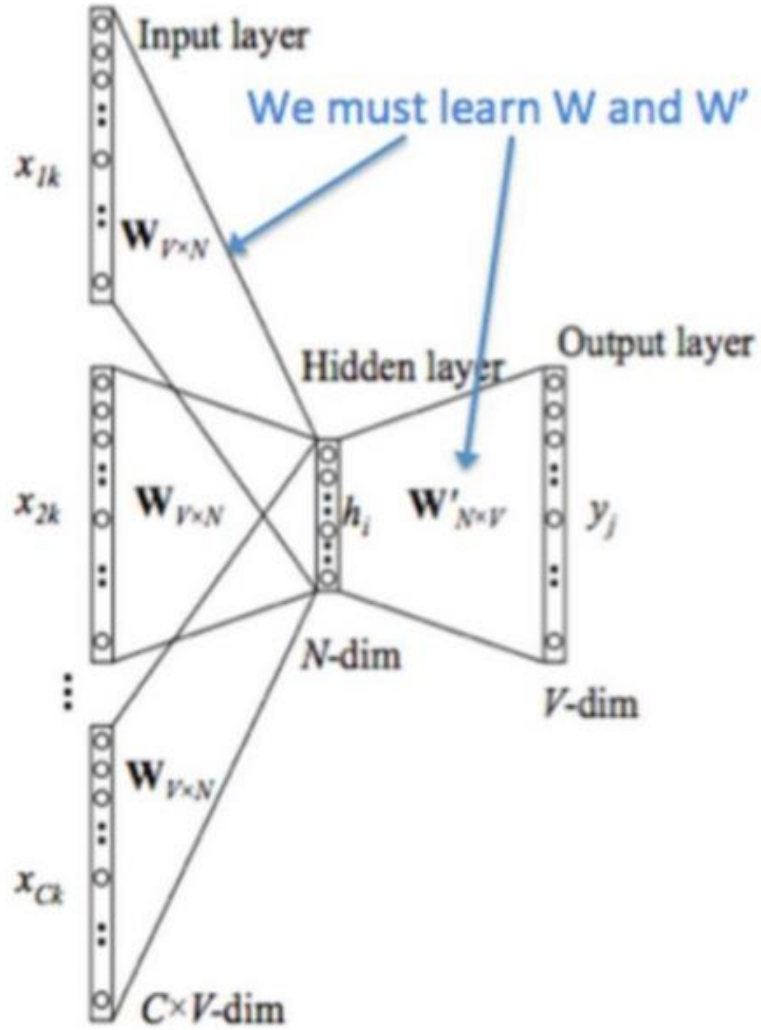
Mô hình CBOW (Continuous BOW)

- Mô hình dự đoán từ hiện tại bằng ngữ cảnh
- Duyệt qua tập dữ liệu lớn với một cửa sổ

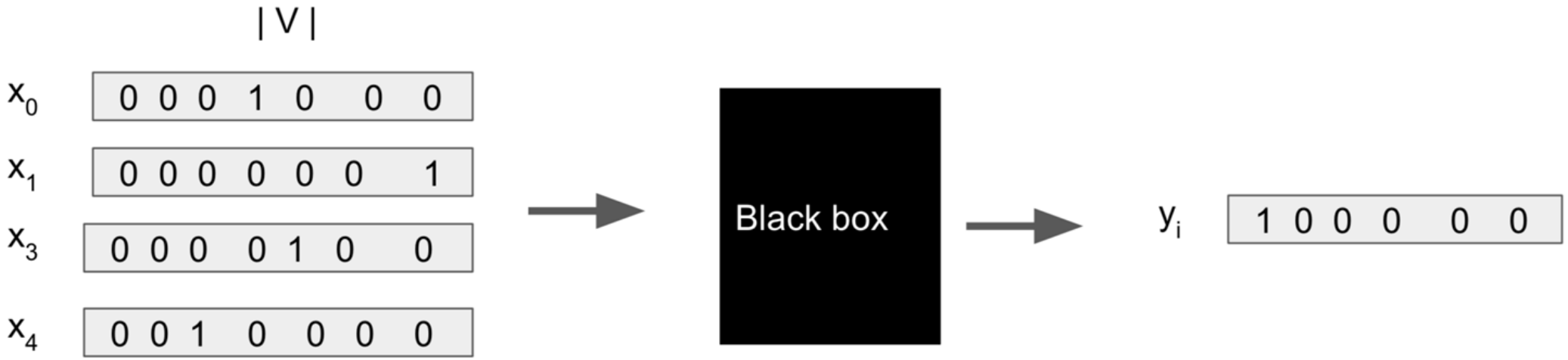
Input : x_0 , x_1 , x_3 , x_4 output : x_2

“ The Cat Chills on a mat ”

x_0 x_1 x_2 x_3 x_4 x_5

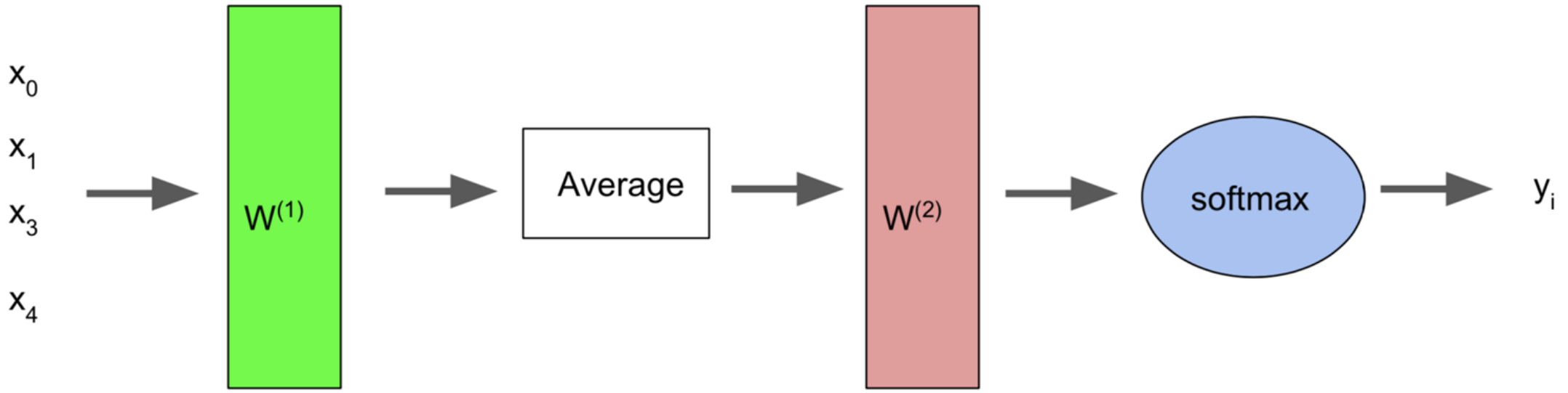


Mô hình CBOW



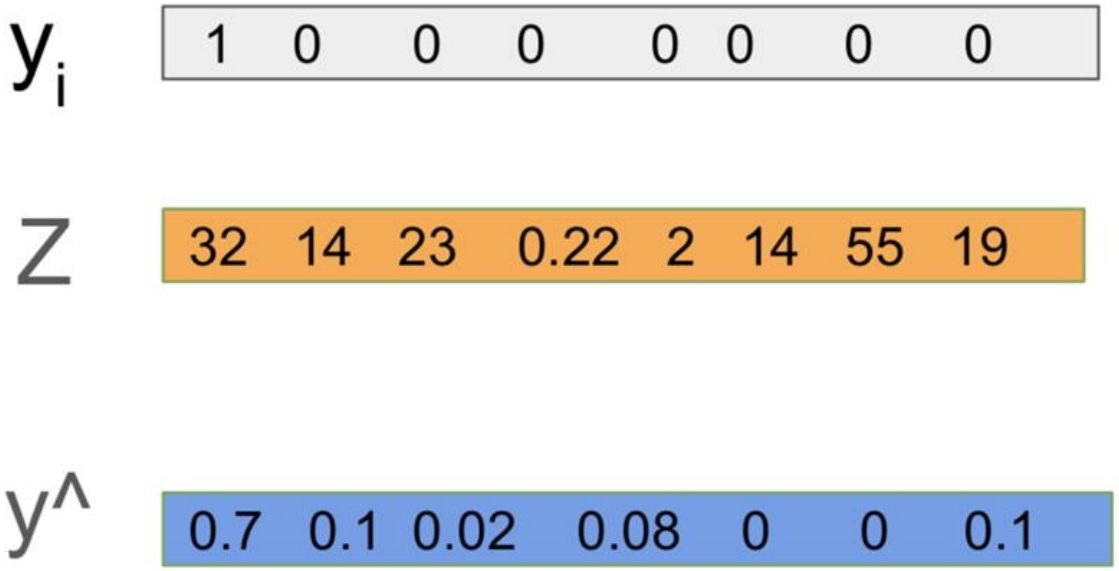
- |V| là kích thước từ điển
- $x_i \in R^{1 \times |V|}$ là biểu diễn one-hot cho 1 từ
- $y_i \in R^{|V| \times 1}$ là biểu diễn one-hot cho từ đúng, ở giữa (từ mong muốn)

Mô hình CBOW



- |V| là kích thước từ điển
- $x_i \in R^{1 \times |V|}$ là biểu diễn one-hot cho 1 từ
- $y_i \in R^{|V| \times 1}$ là biểu diễn one-hot cho từ đúng, ở giữa (từ mong muốn)

Mô hình CBOW – hàm softmax



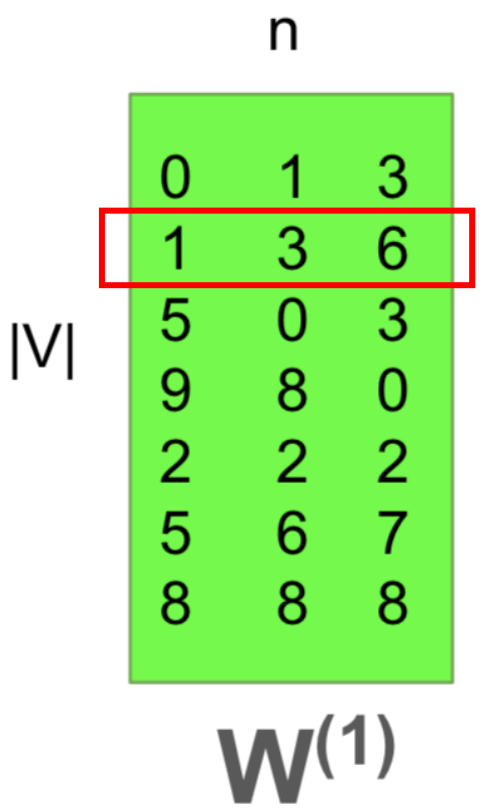
$\hat{y} = \text{softmax}(Z)$

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

$y_i \in R^{|V| \times 1}$ là biểu diễn one-hot cho từ đúng, ở giữa (từ mong muốn)

Vector của mỗi từ

- Sau quá trình huấn luyện, mỗi hàng trong ma trận $W^{(1)}$ chứa giá trị của các từ.
- Các vector biểu diễn từ biểu diễn tốt hơn về mặt cú pháp và ngữ nghĩa
- Cho kết quả tốt trong các bài toán NLP



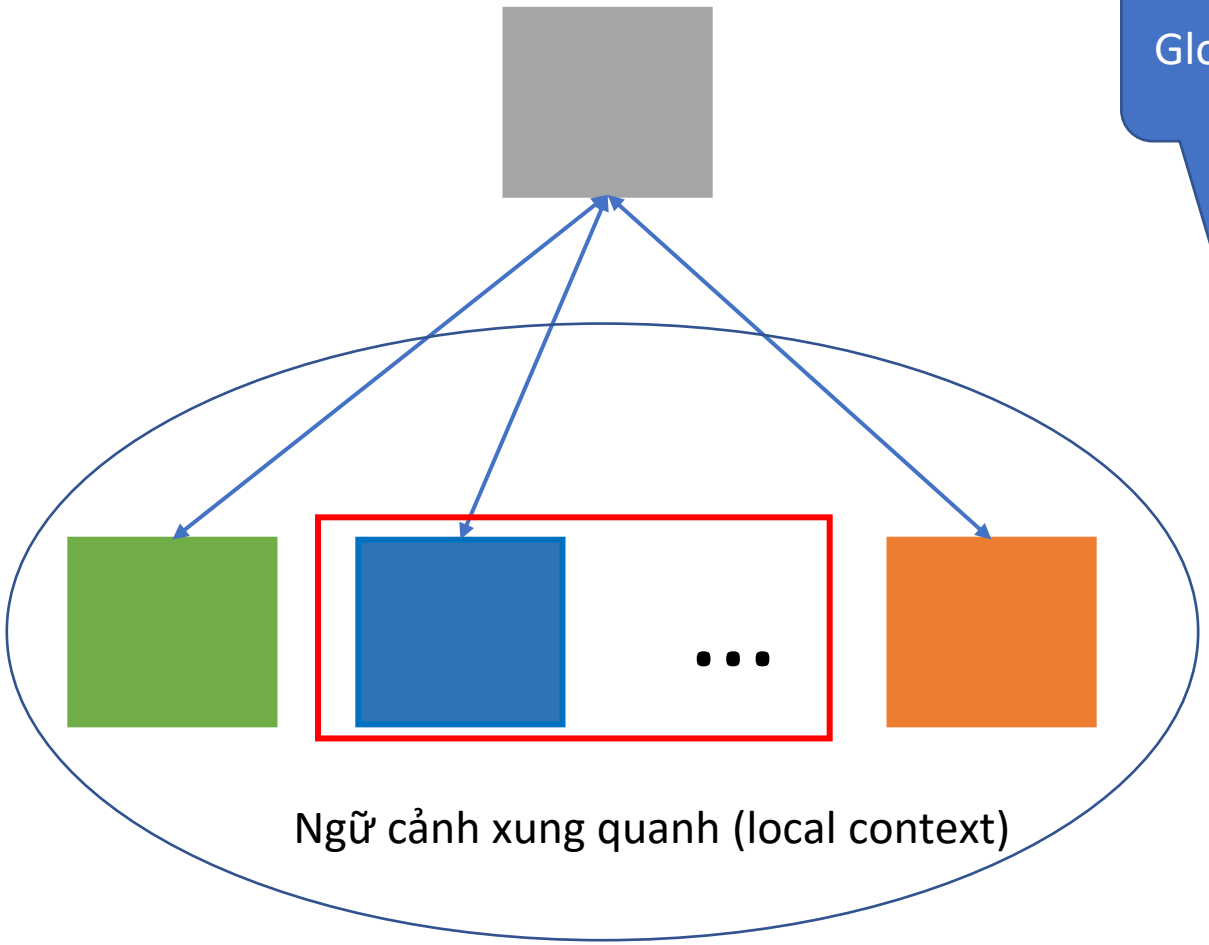
MỘT SỐ MÔ HÌNH BIỂU DIỄN KHÁC

Glove

- Glove (Global Vectors for Word Representation)
- Dựa trên hai giả định
  - Global context (global matrix factorization)
    - Thống kê đồng xuất hiện của các từ trên tập dữ liệu
    - Phân tích dựa trên ma trận
  - Local context
    - Sử dụng cửa sổ trượt
- Qúa trình huấn luyện chỉ dựa trên một ma trận đồng xuất hiện sử dụng phân tích ma trận (matrix factorization)
- Ý tưởng: Glove tạo ra các biểu diễn từ dựa trên sự đồng xuất hiện của hai từ *i* và *j*

https://nlp.stanford.edu/projects/glove/

Word2Vec vs. Glove



Ngữ cảnh xung quanh (local context)

Global context

Local context

$X =$

	I	*like*	*enjoy*	*deep*	*learning*	*NLP*	*flying*	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Ma trận đồng xuất hiện

https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010

Xác suất và tỷ lệ

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

- *P(k|ice)*: xác suất đồng xuất hiện của "*k*" và "*ice*"
  - Số lần "*k*" và "*ice*" cùng xuất hiện chia cho số lần "*ice*" xuất hiện
- Cho trước 2 từ, nếu từ thứ 3 ("*k*")
  - Giống với "ice" nhưng khác với "steam" (solid), xác suất lớn hơn 1
  - Giống với "steam", khác với "ice" (gas), xác suất rất nhỏ
  - Giống hoặc khác cả 2 từ, xác suất gần 1.

Một số vấn đề tính toán

- Không có công tính tính cho ***F(i, j, k) = P_ik/P_jk***
- Vector của từ dạng nhiều chiều, nhưng ***P_ik/P_jk*** là một vector vô hướng
- Hàm ***F(i, j, k)*** gồm ba thành phần (3 từ)
  - Hàm mục tiêu trên 3 thành phần không phù hợp
  - Cần được giảm xuống 2 thành phần

Một số ký hiệu

- ***X***: là ma trận đồng xuất hiện của các từ
- X_{ij}: số lần xuất hiện của từ ***j*** trong ngữ cảnh với từ ***I***
- $X_i = \sum_k X_{ik}$ là số lần xuất hiện của bất kỳ từ nào trong ngữ cảnh với từ ***i***
- $P_{ij} = P(i|j) = X_{ij} / X_i$

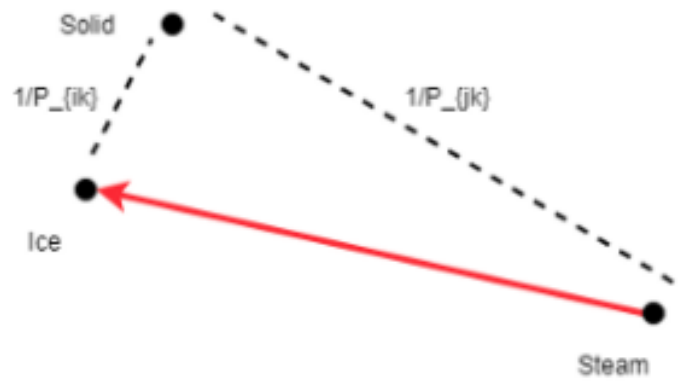
Định nghĩa hàm chi phí

- Hàm chi phí biểu diễn tỷ lệ xác suất đồng xuất hiện của từ

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}},$$

- Sử dụng tính chất tuyến tính của vector

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}.$$



Định nghĩa hàm chi phí

- Sử dụng **dot product**

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}},$$

- Biến đổi vế phải dụng ràng buộc đồng hình (homomorphism)

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}, \quad \text{với} \quad F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}.$$

Định nghĩa hàm chi phí

- Sử dụng hàm mũ Exp (hoặc đảo ngược là log)

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i) \,.$$

- Chỉ phụ thuộc vào *k*, rồi thêm hai bias vectors

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) \,.$$

Hàm chi phí trọng số

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 ,$$

1. f(0) = 0. Nếu f là hàm liên tục, nó sẽ bị triệt tiêu khi $x \to 0$ đủ nhanh mà $\lim_{x \to 0} f(x)\log^2 x$ là hữu hạn
2. f(x) cần không giảm sao cho các đồng xuất hiện hiếm không quá lớn.
3. f(x) tương đối nhỏ cho các giá trị lớn của x, để tần suất đồng suất hiện không quá lớn

Glove trong tính độ tương tự

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

rana

eleutherodactylus

https://www.slideshare.net/JiHyunPark18/understanding-glove?from_action=save

Ưu và nhược điểm

- Ưu điểm
  - Quá trình huấn luyện nhanh, do chỉ sử dụng phân tích ma trận
  - Chất lượng ổn định trên các tập dữ liệu lớn
  - Kết quả tốt ngay cả với tập dữ liệu và vector nhỏ
  - Dừng sớm (early stopping)
    - Có thể dừng qúa trình huấn luyện sớm khi độ cải thiệt không tăng
- Nhược điểm
  - Sử dụng nhiều bộ nhớ do phải lưu trữ ma trận $\boldsymbol{X}$
  - Có thể bị ảnh hưởng bởi quá trình khởi tạo “learning rate”

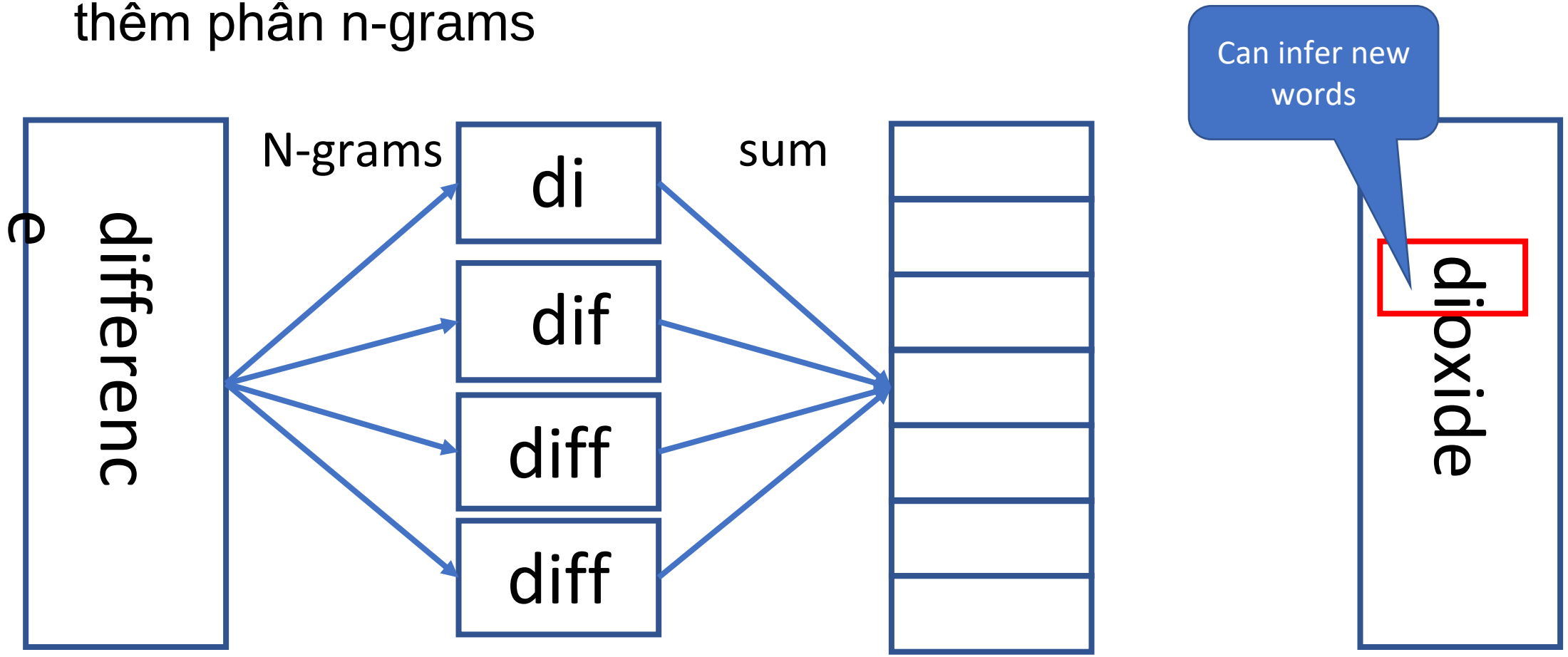
FastText

- FastText
  - Mở rộng từ word2vec
  - Phát triển bởi Facebook
  - Hỗ trợ nhiều ngôn ngữ
  - Thêm thuật ngữ từ nhỏ (sub-word)
    - Dựa trên mô hình n-grams
    - “difference” được huấn luyện bởi “di”, “dif”, “diff”
    - Cho phép học biểu diễn của các từ ngắn
    - Cho phép biểu diễn tiền tố (prefixes) và hậu tố (suffixes) của một từ
  - FastText cho kết quả tốt khi biểu diễn những từ hiếm (rare words)
    - Do sử dụng sub-word)

https://github.com/facebookresearch/fastText/releases

Biểu diễn từ của FastText

- Sử dụng mô hình giống như skip-gram của wordvec, nhưng thêm phần n-grams



Ưu và nhược điểm

- Ưu điểm
  - Biểu diễn tốt với các từ hiếm do sử dụng n-grams
- Nhược điểm
  - Sử dụng nhiều bộ nhớ do phải lưu trữ trên ký tự

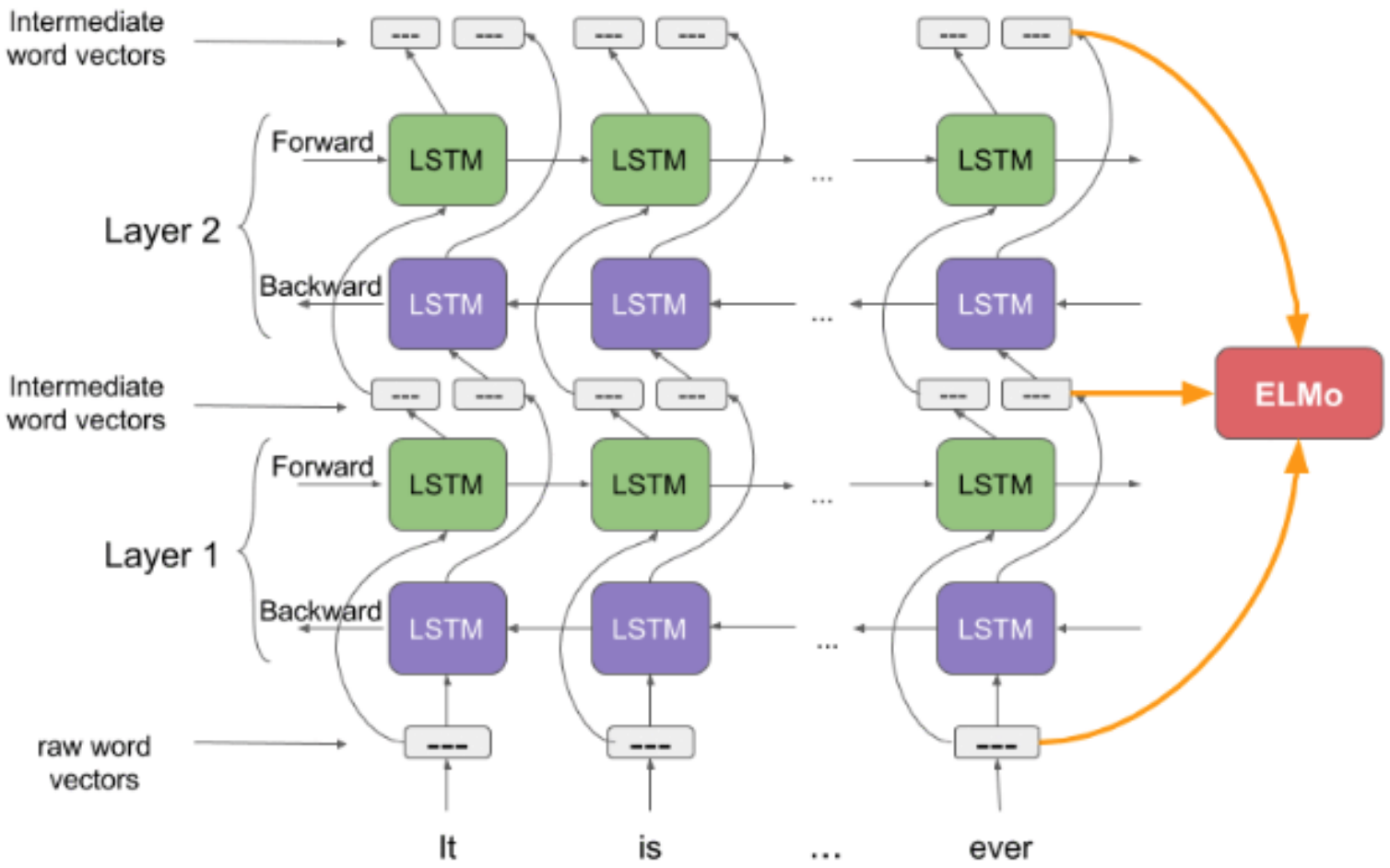
ELMo

- Embedding from Language Model

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

https://arxiv.org/pdf/1802.05365.pdf

Kiến trúc của ELMo



https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/

Kiến trúc của ELMo

- Dùng character-level CNN để biểu diễn từ (raw word vectors)
- Sử dụng hai tầng biLM (bidirectional LM)
- Sự kết hợp theo chiều tiến và lùi của một từ tạo thành vector của tầng giữa (intermediate word vectors)
- Tầng biLM thứ 2 dùng vectors ở tầng intermediate word vectors
- Biểu diễn cuối cùng của một từ là vector trọng số của:
  - Biểu diễn từ ban đầu
  - Và vectors của 2 tầng biLM

https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/

Ưu và nhược điểm

- Ưu điểm
  - Biểu diễn ngữ nghĩa tốt hơn so với word2vec và Glove
    - Hai từ giống nhau có vector khác nhau dựa vào ngữ cảnh

      I ***read*** the book yesterday vs. Can you ***read*** the letter now?

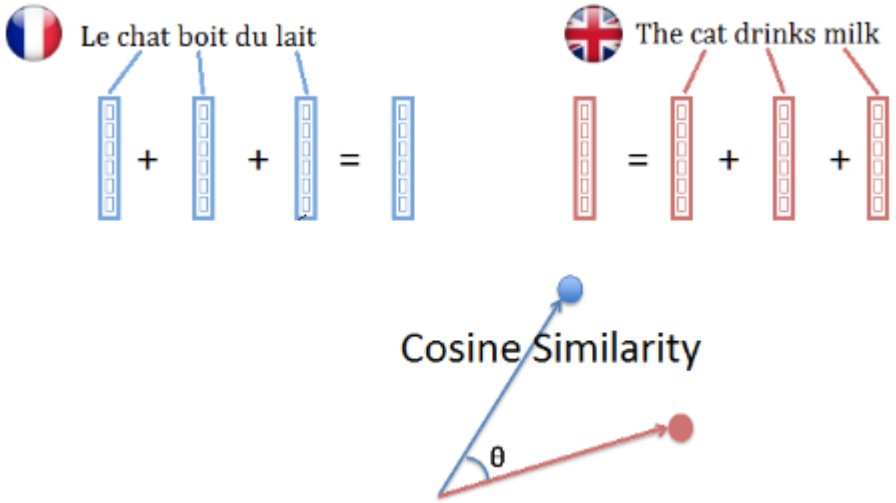
    - Xử lý tốt vấn đề từ hiếm, do biểu diễn trên ký tự
- Nhược điểm
  - Sử dụng nhiều bộ nhớ do phải lưu trữ trên ký tự

Xây dựng mô hình biểu diễn từ

- Thu thập dữ liệu đủ lớn (wikipedia)
- Lựa chọn mô hình phù hợp với bài toán
- Huấn luyện mô hình
  - Thường cần GPU
- Sử dụng mô hình vào các bài toán cụ thể

Một số ứng dụng trong phát hiện đạo văn

- Phát hiện đạo văn
  - Một ngôn ngữ (PlagiarismDetection with a WORD2VEC)
  - 2 ngôn ngữ (Using Word Embedding for Cross-Languag Plagiarism Detection)



Tổng kết

- Giới thiệu bài toán phát hiện đạo văn
- Nhắc lại một số kiến thức về vector hoá và word2vec
- Một số phương pháp biểu diễn khác
  - Glove
  - FastText
  - ELMo
- Xây dựng mô hình biểu diễn từ
- Ứng dụng biểu diễn từ cho bài toán phát hiện đạo văn

THỰC HÀNH

Q&A

Thank you!