VietAI

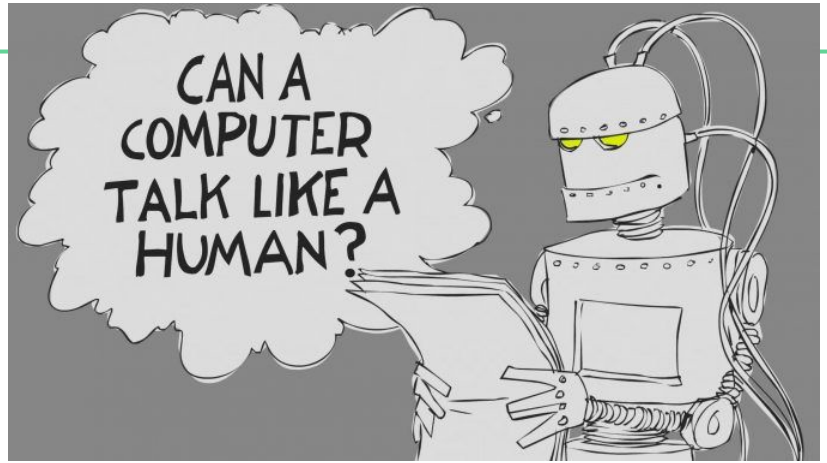
Bài 11: Deep Learning trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP): Biểu diễn từ

Tuần 6A
23-9-2019

Nội dung chính

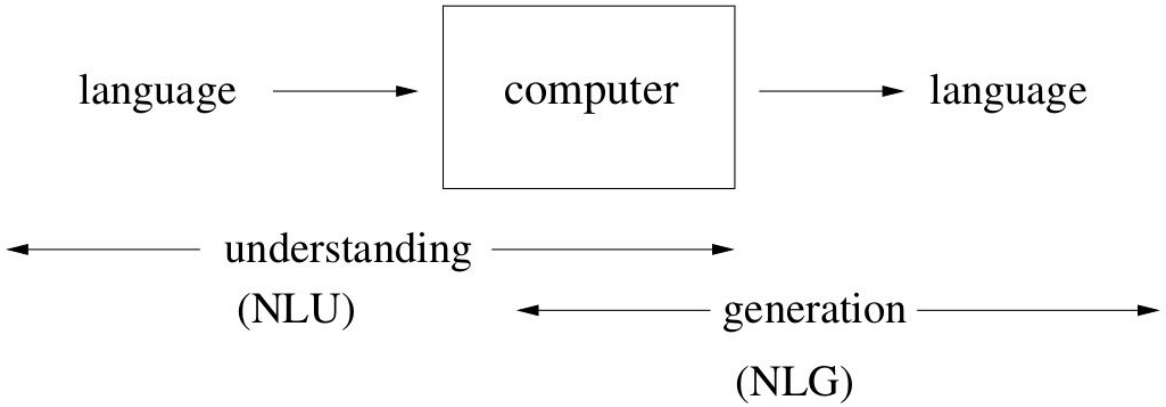
1. Giới thiệu lĩnh vực Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)
2. Ứng dụng Deep Learning trong các bài toán NLP
3. Biểu diễn từ (word representation) bằng Mô hình Word2Vec

1. Giới thiệu lĩnh vực Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

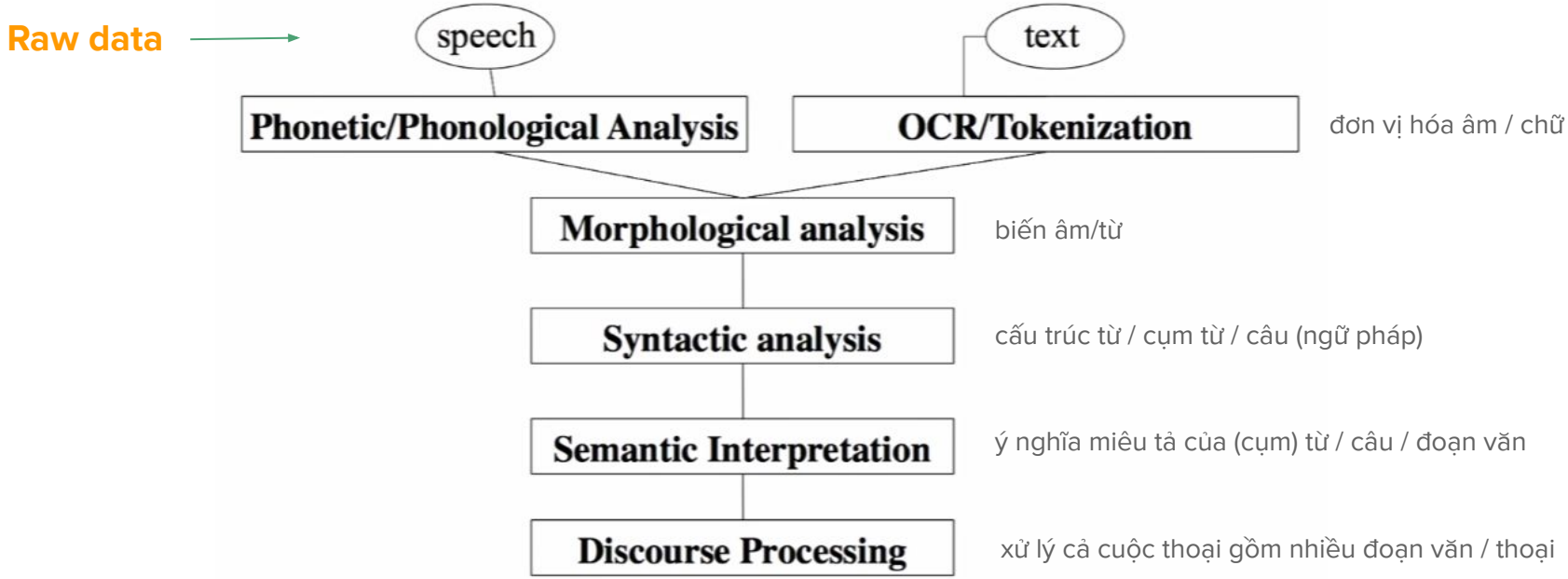


Tổng quan

- ***Xử lý ngôn ngữ tự nhiên*** (***Natural Language Processing -*** *NLP*) là ngành nghiên cứu kết hợp giữa khoa học máy tính (CS), trí tuệ nhân tạo (AI) và ngôn ngữ học (Linguistics) -> *xử lý dữ liệu dạng text*
- Mục tiêu: máy tính có thể "hiểu" và “tạo” được ngôn ngữ tự nhiên của con người



Tổng quan



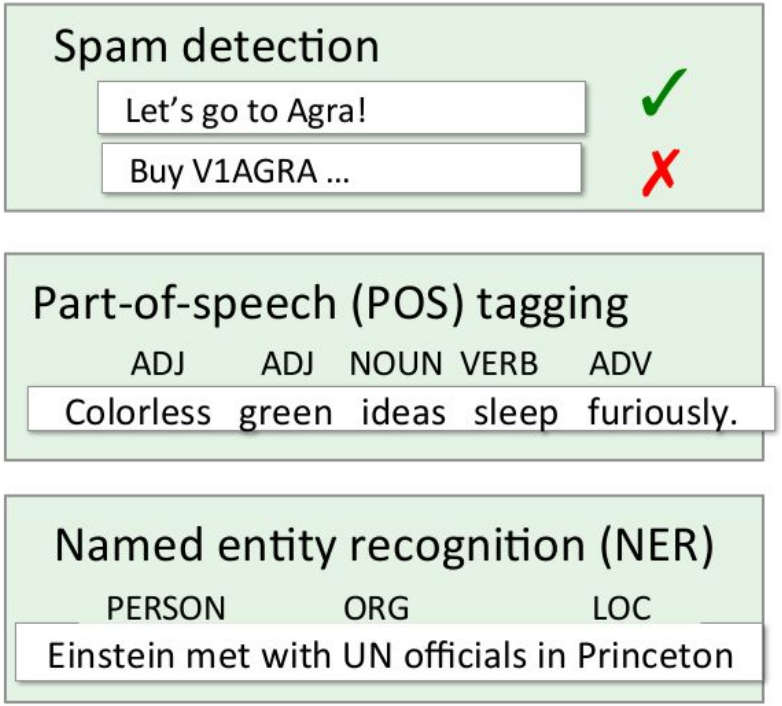
Các mức độ phân tích để “hiểu” ngôn ngữ

Những sự phức tạp trong các bài toán NLP

- Tính nhập nhằng (ambiguity) *"Chiếc cúp không vừa với vali vì **nó** quá lớn."*
- Vấn đề trong segmentation *"Tốc độ truyền thông tin ..."*
- Ngôn ngữ không theo chuẩn *"M0ther ui, hum n4i con hk zia, k0n f4i h0k th3m"*
- Thành ngữ *"ra ngô ra khoai"*
- Phụ thuộc vào ngữ cảnh (context) và kiến thức ở thế giới thực.

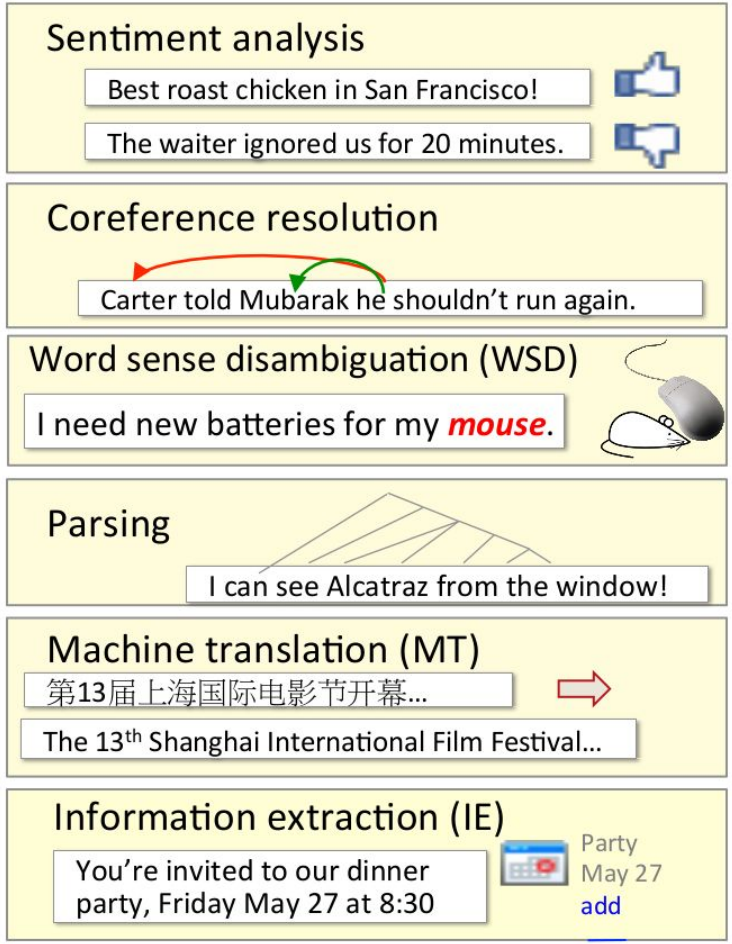
Một số bài toán trong NLP

- Kiểm tra lỗi chính tả - Spell checking
- Nhận dạng thư rác - Spam detection
- Gán nhãn từ loại - Part-of-speech tagging
- Nhận dạng thực thể có tên - Named Entity Recognition (NER)
- Tìm kiếm từ khóa
- Tìm từ đồng nghĩa



Một số bài toán trong NLP

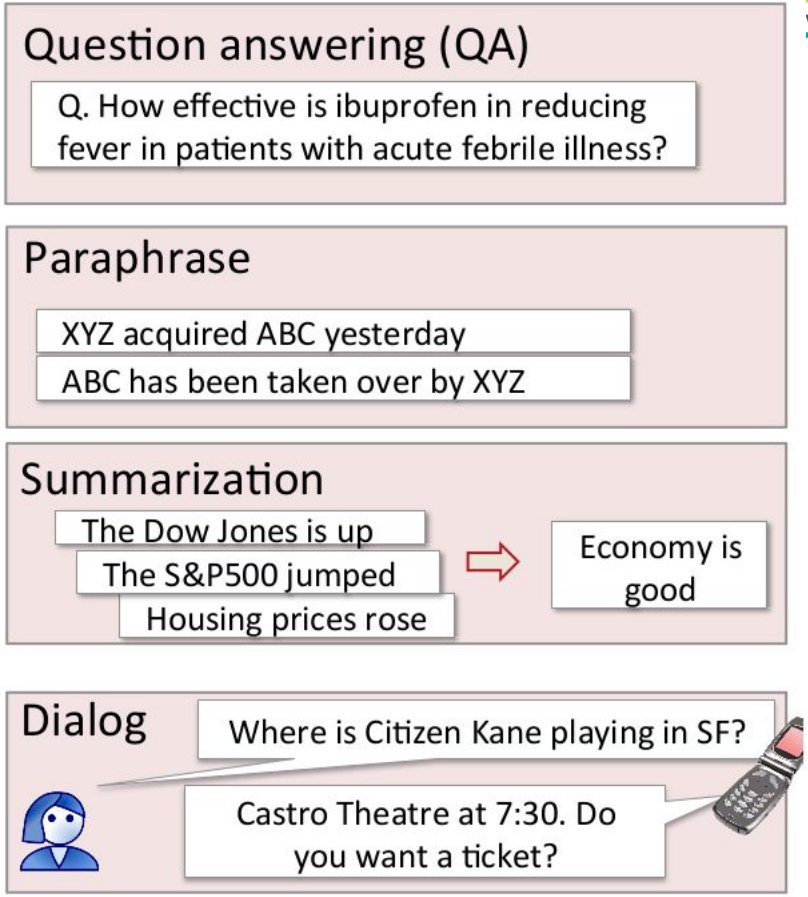
- Phân tích cảm xúc - Sentiment Analysis / Opinion Mining
- Coreference resolution
- Xác định ý nghĩa được sử dụng của một từ nhiều nghĩa - Word-sense disambiguation
- Parsing
- Dịch máy - Machine Translation
- Trích xuất thông tin - Information Extraction



Neural Network in NLP

Xử lý ngôn ngữ tự nhiên (NLP) - Ứng dụng

- Hỏi đáp - Question Answering
- Viết lại - Paraphrase
- Tóm tắt văn bản - Text Summarization
- Hệ thống đối thoại - Spoken Dialog System



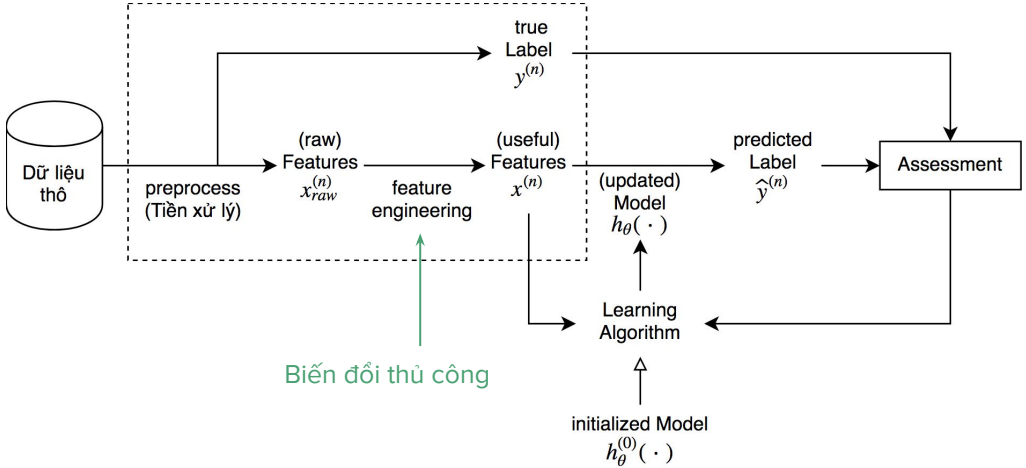
2. Ứng dụng Deep Learning trong các bài toán NLP

Sơ lược

- Deep NLP = Deep Learning + NLP
- Sử dụng các mô hình Deep Learning để
  - học các biểu diễn (representation) có ý nghĩa learning
  - làm mô hình chính giải quyết các bài toán trong NLP.
- Đem lại nhiều tiến bộ vượt bậc trong những năm gần đây trên các phương diện khác nhau:
  - Syntax, Semantics
  - Part-of-speech, Named Entity Recognition, Parsing
  - Machine Translation, Sentiment Analysis
  - Question Answering, Dialogue Agents (i.e. chatbot)

Biểu diễn từ i.e. Tạo semantic cho từ (word token)

Biến đổi đặc trưng thủ công (hand-designed) để tạo semantic cho từ (word token)



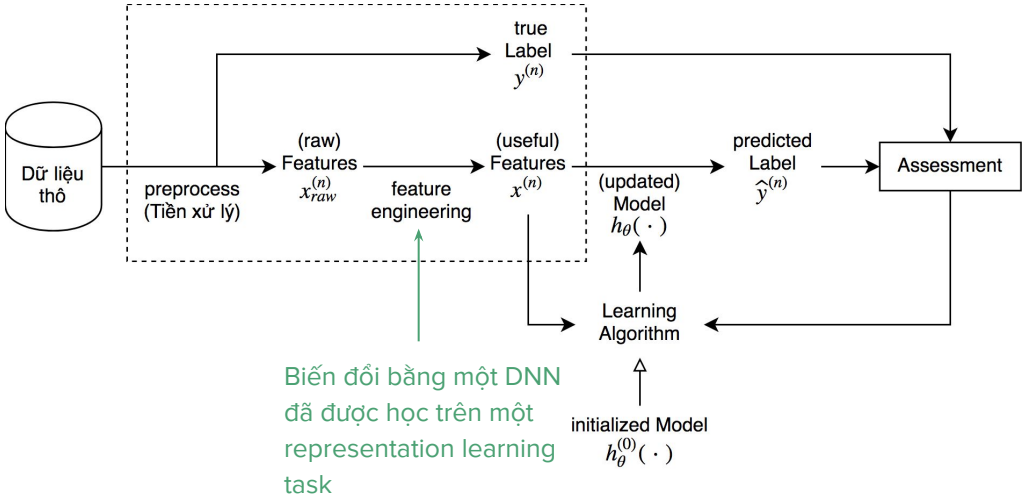
Feature	NER
Current Word	✓
Previous Word	✓
Next Word	✓
Current Word Character n-gram	all
Current POS Tag	✓
Surrounding POS Tag Sequence	✓
Current Word Shape	✓
Surrounding Word Shape Sequence	✓
Presence of Word in Left Window	size 4
Presence of Word in Right Window	size 4

Một số vd. các hand-designed features được sử dụng để giải quyết bài toán nhận dạng thực thể - NER - đối với địa danh và tên tổ chức (Finkel, 2010)

Biểu diễn từ i.e. Tạo semantic cho từ (word token)

Biến đổi đặc trưng bằng một Mô hình DNN *đã được học trên một task khác* - cụ thể: *word-level* ***representation learning*** task - để tạo semantic cho từ (word token)

- Word2Vec
- GLoVE
- ELMo
- BERT



Biểu diễn từ i.e. Tạo semantic cho từ (word token)

- Biểu diễn từ dưới dạng vector -> word-to-vector -> “word2vec”
- Tập hợp tất cả các vector biểu diễn từ được gọi là **Word Embedding**

$$\text{august} = \begin{bmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.109 \\ 0.107 \\ -0.542 \\ 0.349 \end{bmatrix}$$

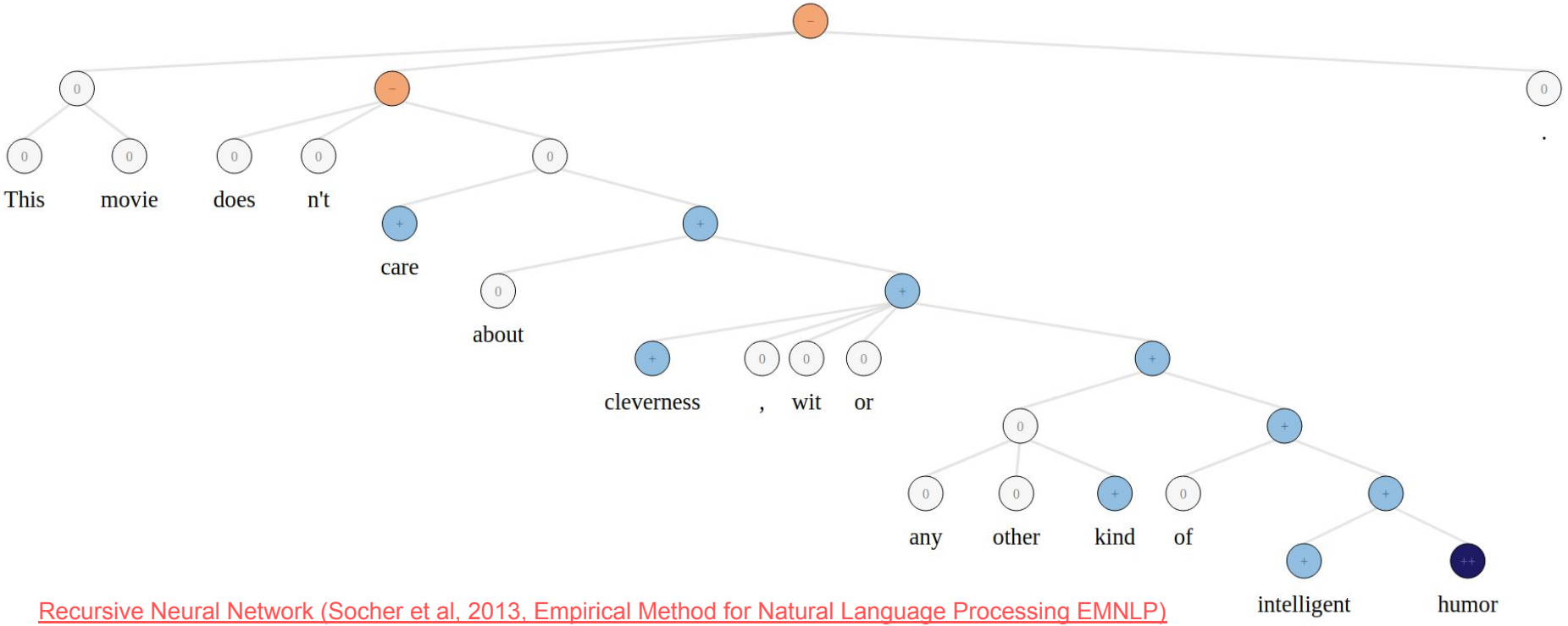
Biểu diễn từ i.e. Tạo semantic cho từ (word token)

Ví dụ

Vd. Đây là top 7 từ "gần nhất" (đo bằng cosine similarity) với từ “frog” sau khi các vector được học với mô hình *GloVe (Global Vector)*:

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

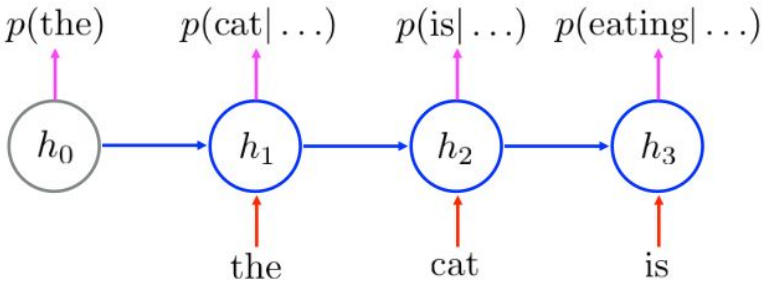
Phân tích cảm xúc bằng Mô hình Recursive NN



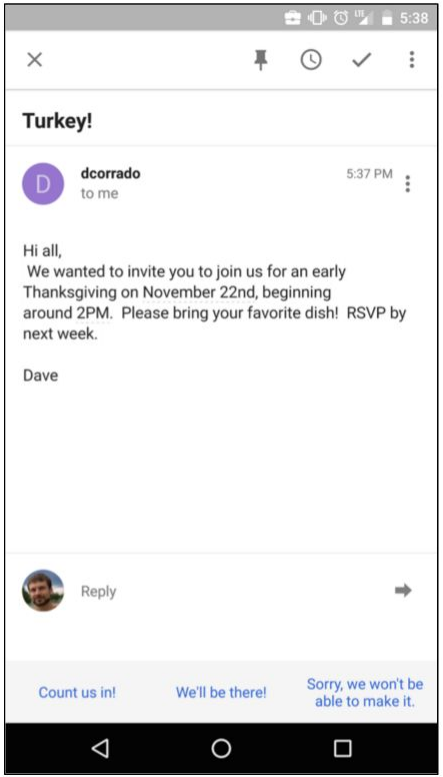
Recursive Neural Network (Socher et al, 2013, Empirical Method for Natural Language Processing EMNLP)

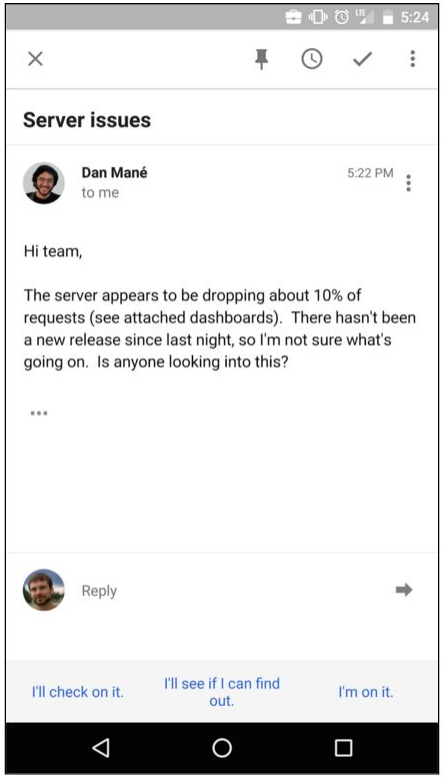
Xây dựng Language Model bằng RNN

- Ứng dụng: Dialogue Agent và Response Generation
  - Vd. Chức năng *smart reply* trên Google Inbox/GMail.



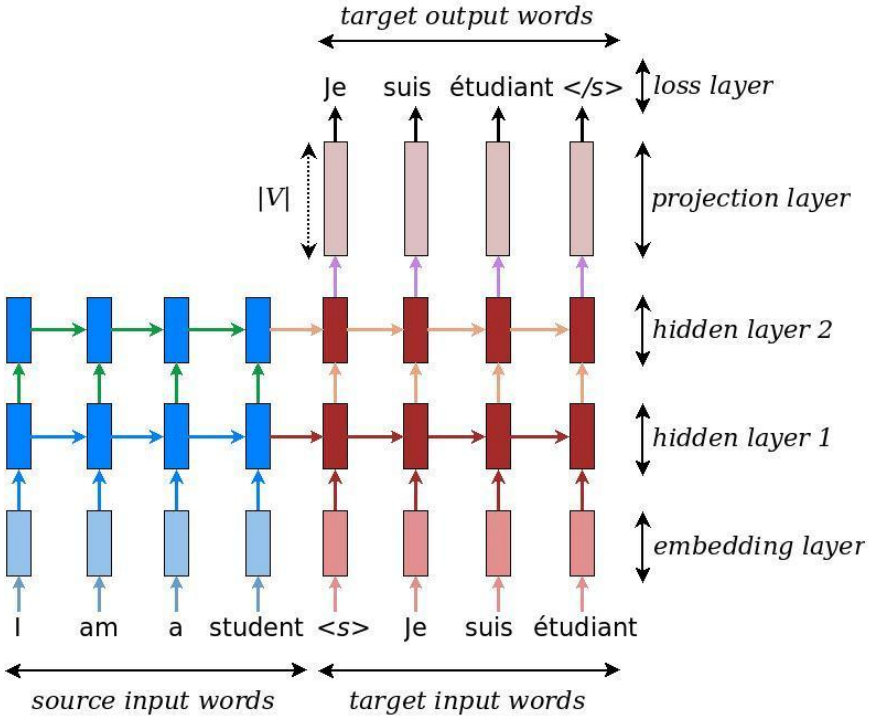
Neural Language Model





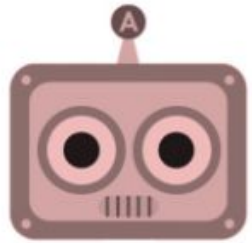
Dịch máy tự động bằng Mô hình Seq2Seq

- Các phương pháp truyền thống tiếp cận theo nhiều phương diện khác nhau (trực tiếp, syntactic, semantic)
- Các phương pháp dịch máy truyền thống thường rất lớn và phức tạp.
- Sử dụng Seq2Seq - một loại kiến trúc Recurrent Neural Network (RNN) đặc thù - để encode câu input thành một vector rồi decode vector đó thành câu output.



Question Answering

TASK	**Question Answering**
DATASET	Natural Questions
EXAMPLES	*Who wrote the book the origin of species?* **Correct answer**: *Charles Darwin* **Model answer**: `Charles Darwin` *What is the largest state in the U.S. by land mass?* **Correct answer**: *Alaska* **Model answer**: `California`
PERFORMANCE	



Text summarization

TASK	**Summarization**: summarize news articles
DATASET	CNN and Daily Mail dataset
EXAMPLE TEXT (TRUNCATED FOR BREVITY)	*Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.* *Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.* *...*
REFERENCE SUMMARY	*Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.*
SUMMARY (MACHINE-WRITTEN)	`The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.`

VietAI

Biểu diễn từ (Word representation): Mô hình Word2Vec

Word2Vec

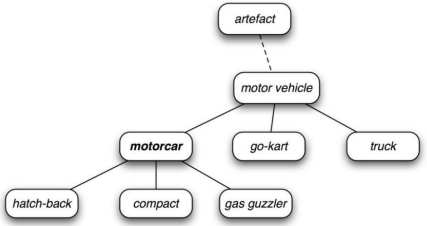
Outline

- Giới thiệu chung về các phương pháp biểu diễn từ
- Giới thiệu chung về Word2Vec
  - Mô hình Continuous Bag-of-words (CBOW)
  - Mô hình Skip-gram
- Cosine Similarity
- Kết quả
- Học vector của các cụm từ

Word2Vec

Giới thiệu chung về các phương pháp biểu diễn từ

- Việc biểu diễn từ thành vector là công việc rất quan trọng để áp dụng các phương pháp Machine Learning.
- Các kĩ thuật thường dùng:
  - Knowledge-Based representation: wordnet,...
  - Corpus-based representation
    - Atomic symbol: one-hot vector
    - Counted-based: TF-iDF, LDA,
    - **Context-based: Word2vec**



Word2Vec

Giới thiệu chung về các phương pháp biểu diễn từ

Ví dụ về Counted-based method

... and the *cute* **kitten** *purred* and then ...
... the *cute furry* **cat** *purred* and *miaowed* ...
... that the *small* **kitten** *miaowed* and she ...
... the *loud furry* **dog** *ran* and *bit* ...

basis vocabulary: *{bit, cute, furry, loud, miaowed, purred, ran, small}*.

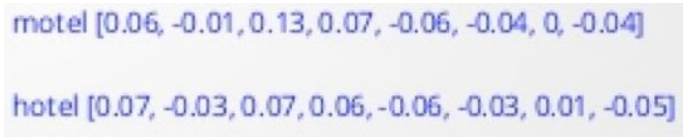
$$\mathbf{kitten} = [0, 1, 0, 0, 1, 1, 0, 1]^\top$$
$$\mathbf{cat} = [0, 1, 1, 0, 1, 0, 0, 0]^\top$$
$$\mathbf{dog} = [1, 0, 1, 1, 0, 0, 1, 0]^\top$$

$$sim(\text{kitten}, \text{cat}) = cosine(\mathbf{kitten}, \mathbf{cat}) \approx 0.58$$
$$sim(\text{kitten}, \text{dog}) = cosine(\mathbf{kitten}, \mathbf{dog}) = 0.00$$
$$sim(\text{cat}, \text{dog}) = cosine(\mathbf{cat}, \mathbf{dog}) \approx 0.29$$

Word2Vec

Giới thiệu chung về các phương pháp biểu diễn từ

Biểu diễn dưới dạng "túi từ" (BOW)



Biểu diễn dưới dạng Distributed Representation

Word2Vec

Giới thiệu chung về các phương pháp biểu diễn từ

"You shall know a word by the company it keeps"

(J. R. Firth, 1957)

I love ***you*** so much

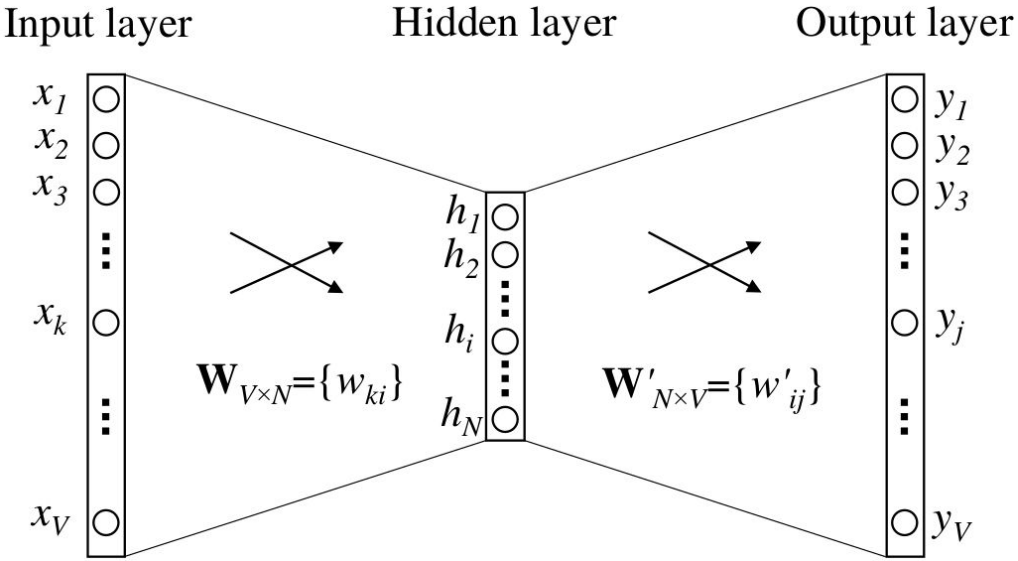
I love ***him*** so much

I love ***her*** so much

Word2Vec

Giới thiệu chung về Word2Vec

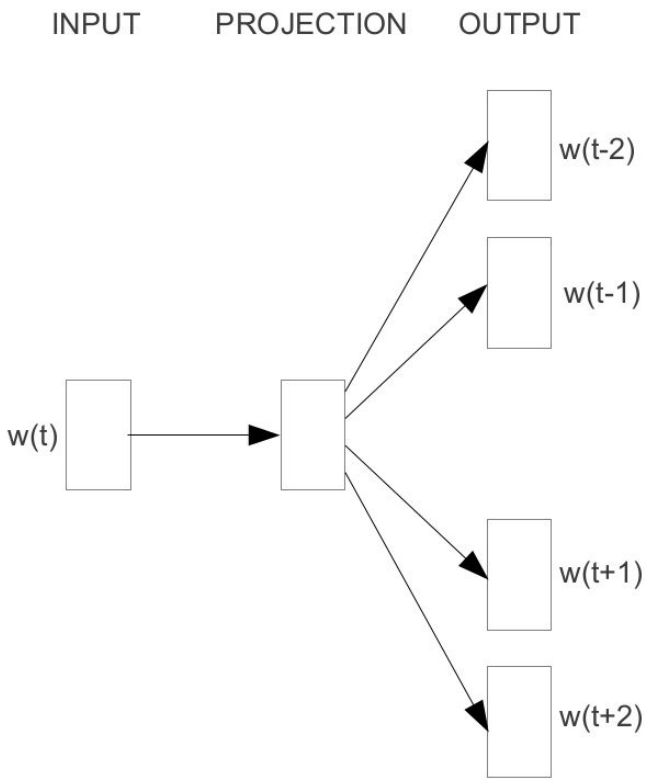
- Word2Vec ra đời năm 2013 bởi Tomas Mikolov và cộng sự từ Google.
- Word2Vec có 2 biến thể: Skip-gram và Continuous Bag-of-words (CBOW)
- Sử dụng 2 kỹ thuật để tăng tốc việc huấn luyện: Hierarchical Softmax và Negative Sampling.



Word2Vec

Mô hình Skip-gram

- Ý tưởng chính của Skip-gram: Dự đoán các từ xung quanh khi cho một từ ở giữa.
- Sử dụng Neural Network với 1 hidden layer và output layer là softmax.

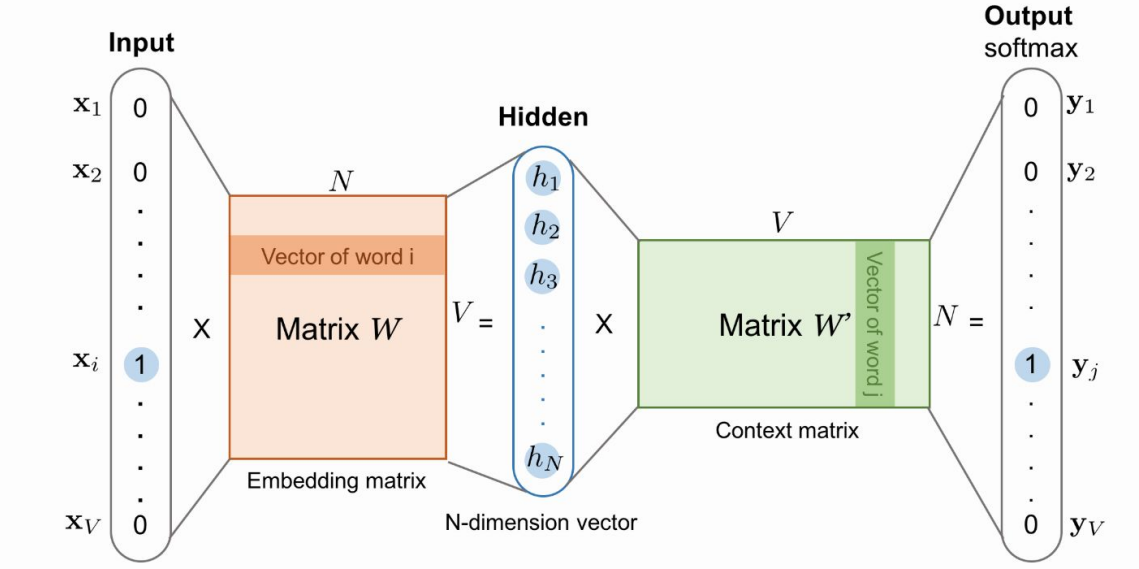


$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p\left(w_{t+j}|w_t\right)$$

Word2Vec

Mô hình Skip-gram

- Vector x : one-hot vector tương ứng với từ w_t
- $h = W^\top x$ việc này tương đương với lấy dòng thứ k tương ứng với từ w_t
- $y = \text{softmax}\,(h; W')$

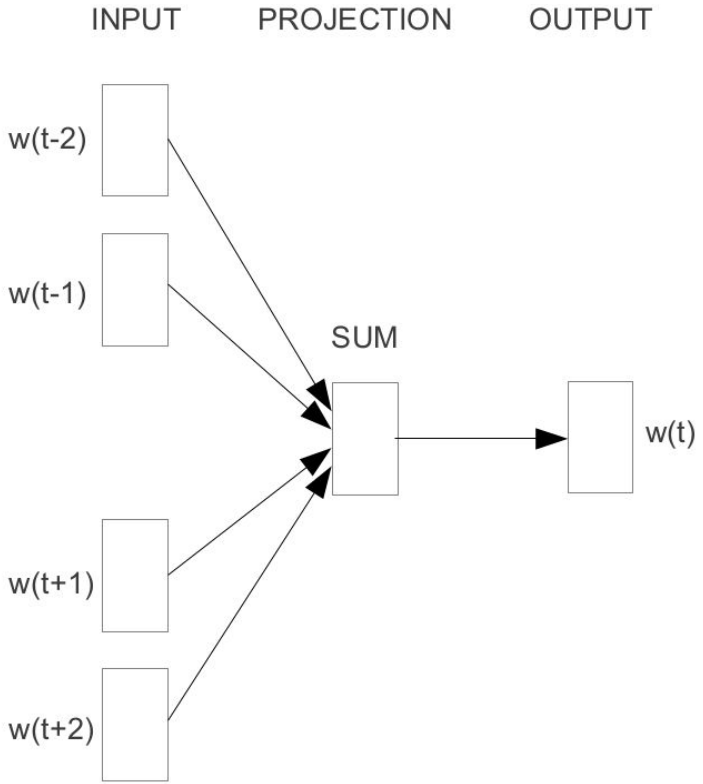


$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p\,(w_{t+j}|w_t)$$

Word2Vec

Mô hình Continuous Bag-of-words

- Ý tưởng chính của Continuous Bag-of-words (CBOW): Dự đoán từ ở giữa khi cho các từ xung quanh.

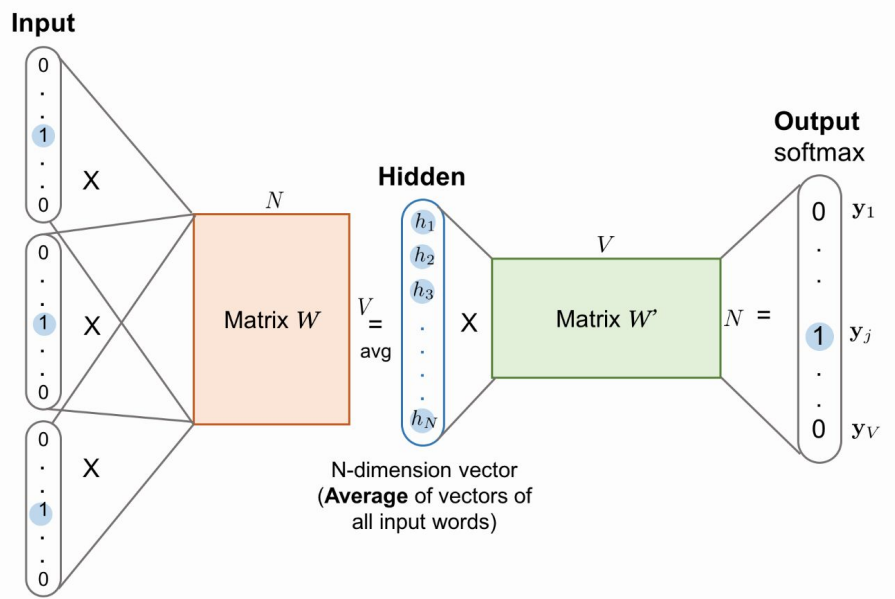


$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p\left(w_t | w_{t-c}, w_{t-c+1}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c-1}, w_{t+c}\right)$$

Word2Vec

Mô hình Continuous Bag-of-words

- Dựa vào các từ đầu vào tính vector h bằng cách lấy tổng hoặc trung bình của các vector tương ứng. Sau đó đưa qua softmax để dự đoán từ ở giữa.
- $h = W^\top (x_1 + x_2 + \cdots + x_C)$
- $y = \text{softmax}\,(h; W')$

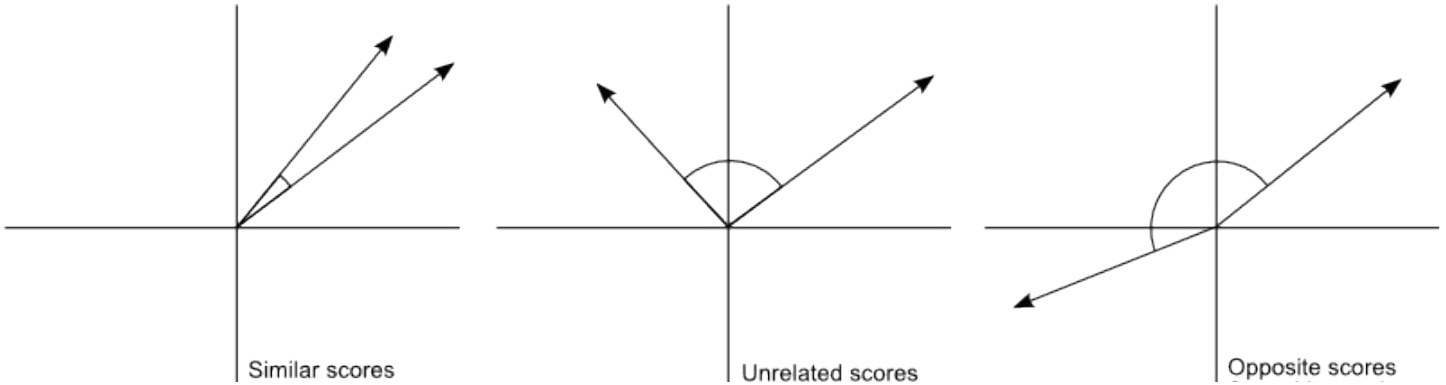


Word2Vec

Cosine Similarity

- Cosine similarity đo độ "tương tự" giữa hai vector theo công thức:

$$\text{cos-sim} = \frac{AB}{\|A\|\|B\|}$$



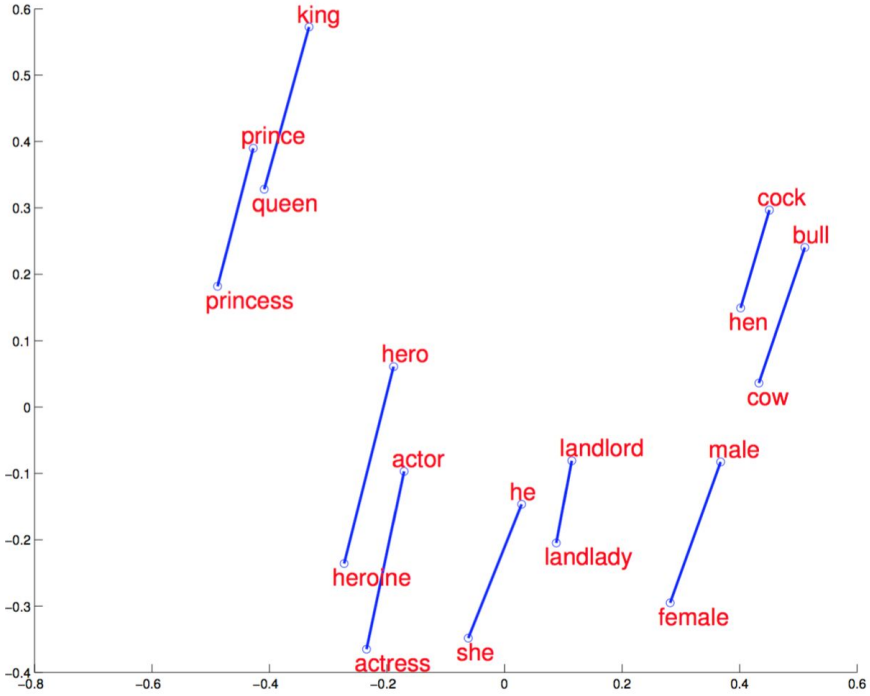
Word2Vec

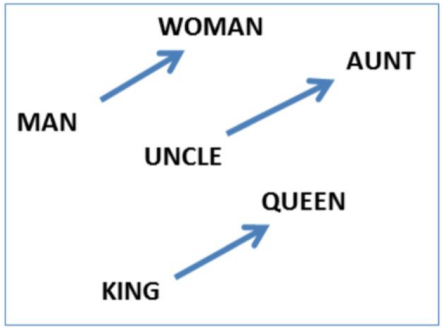
Kết quả

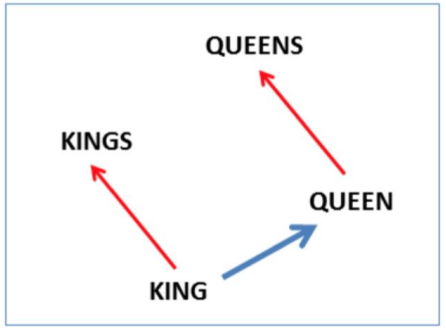
- apples - apple + car = **X** ⇔ apples - apple = **X** - car
- quickly - quick + slow = **Y** ⇔ quickly - quick = **Y** - slow
- King - Man + Woman = **Z** ⇔ King - Man = **Z** - Woman
- Berlin - Germany + France = **T** ⇔ Berlin - Germany = **T** - France

Word2Vec

Kết quả







Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Word2Vec

Kết quả

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Word2Vec

Mở rộng: Học vector biểu diễn các cụm từ

- Để học vector biểu diễn các **cụm từ**, ta thực hiện bước tiền xử lý để gom các từ thường xuyên đứng cạnh nhau trong corpus.

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Take-home Messages

- NLP là ngành phân tích, “hiểu” và tự động sinh ngôn ngữ nói của con người, thể hiện dưới dạng dữ liệu dạng văn bản (text).
- Có nhiều mức độ “hiểu” ngôn ngữ, thể hiện ở những bài toán tương ứng trong NLP. Deep Learning đã và đang có vai trò ngày càng lớn trong việc giải quyết các bài toán này.
- *Biểu diễn từ* là một bài toán NLP nền tảng, biến đổi token từ thành vector chứa nhiều thông tin hơn. Word2Vec, GLoVE, BERT, ... là các phương pháp học biểu diễn từ sử dụng DNN, và dựa trên context (vd. các từ xung quanh) của từ cần biểu diễn
- *Nội dung tiếp theo*: Kiến trúc RNN & Bài toán Language Modelling

Tài liệu tham khảo

1. Lecture 1-3 - *Natural Language Processing with Deep Learning - CS224N* - Stanford University.
2. Word2Vec
   a. Mikolov et al., *Distributed Representations of Words and Phrases and their Compositionality*, NIPS Workshop 2013.
   b. Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, ICLR 2013.
   c. Code: https://code.google.com/archive/p/word2vec/, https://github.com/tmikolov/word2vec