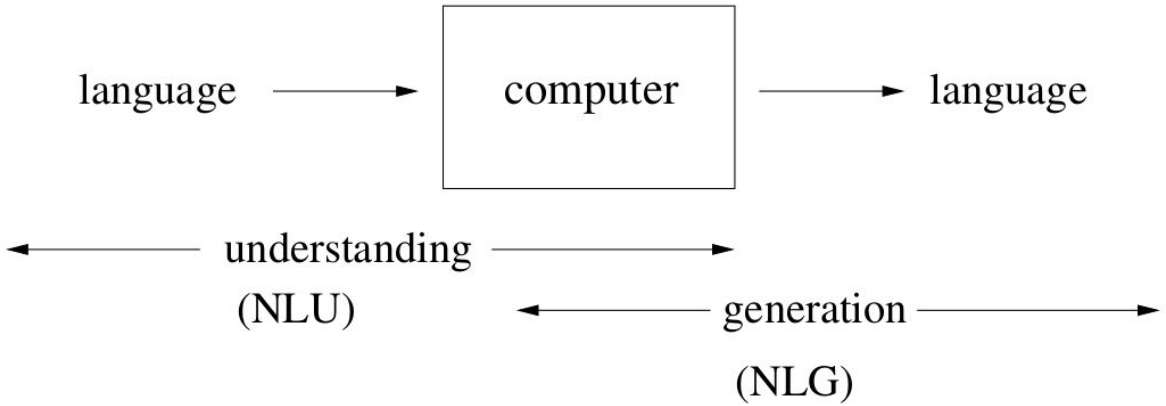
Deep Learning in NLP & Word Representation

VietAI Teaching Team

1 Xử lý ngôn ngữ tự nhiên (NLP)

- ***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).
- Mục tiêu: máy tính có thể "hiểu" được ngôn ngữ tự nhiên của con người.



Nội dung

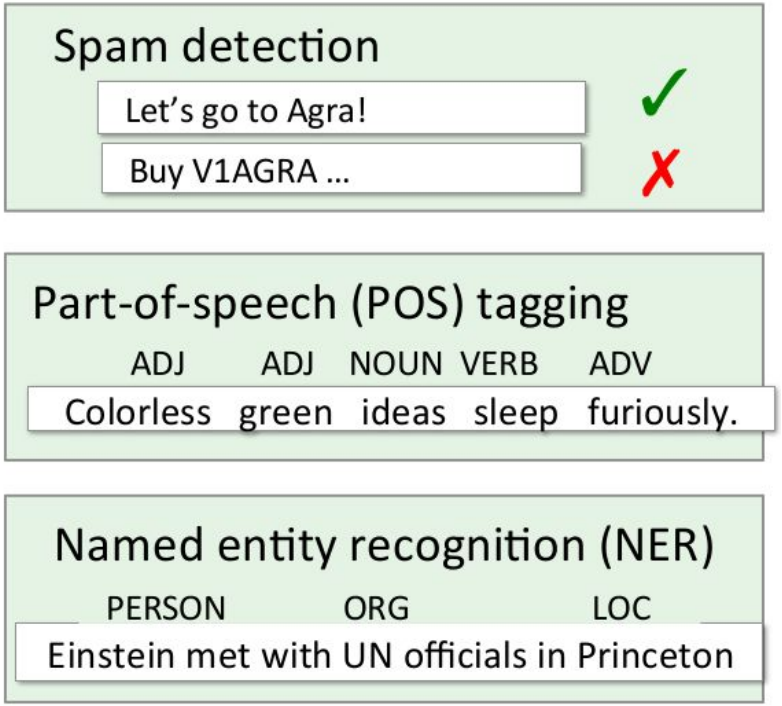
1. Xử lý ngôn ngữ tự nhiên (NLP)
2. Ứng dụng của Deep Neural Network trong NLP
3. Word2Vec

1 Xử lý ngôn ngữ tự nhiên - Khó khăn

- Tính nhập nhằng (ambiguity) *"Ông già đi nhanh quá!"*
- 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.

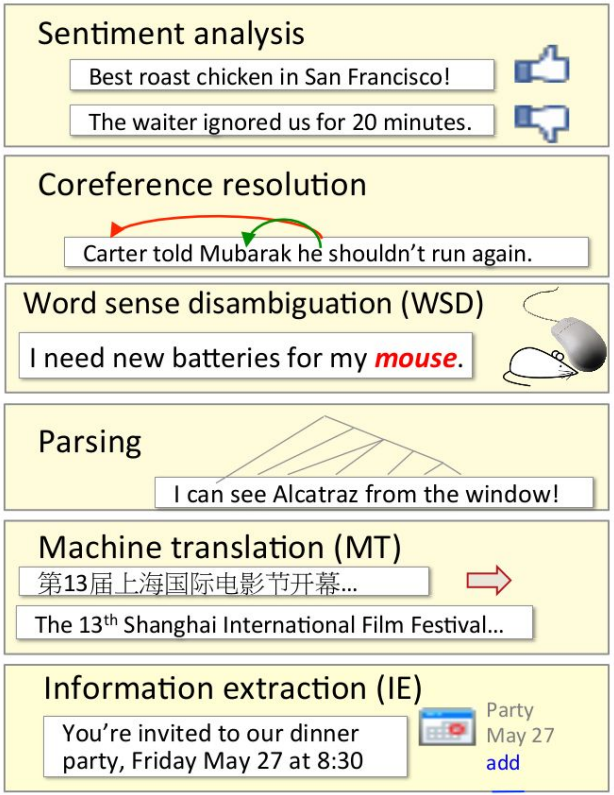
1 Xử lý ngôn ngữ tự nhiên - Ứng dụng

- 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)
- Tìm kiếm từ khóa
- Tìm từ đồng nghĩa



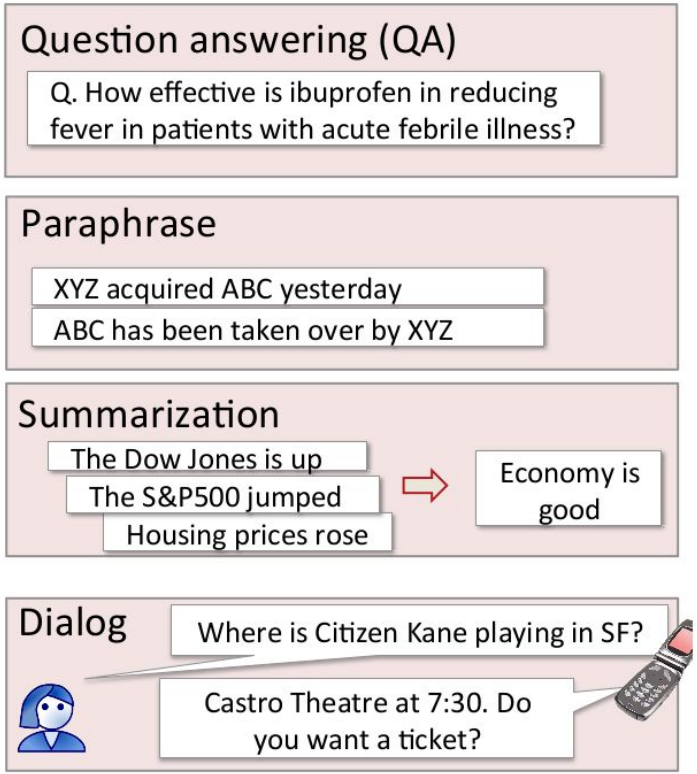
1 Xử lý ngôn ngữ tự nhiên - Ứng dụng

- Sentiment Analysis - Opinion Mining
- Coreference resolution
- Word sense disambiguation
- Parsing
- Dịch máy (Machine Translation)
- Truy xuất thông tin (Information Extraction)



1 Xử lý ngôn ngữ tự nhiên - Ứng dụng

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



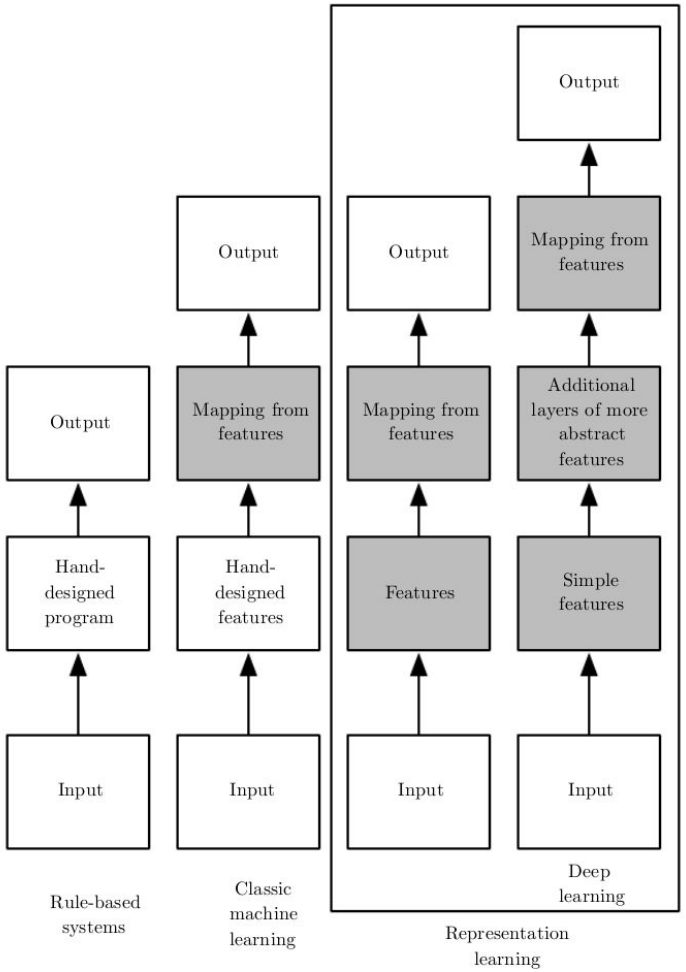
2 Deep Learning in NLP

- Deep Learning (Học sâu) là một ngành con của Machine Learning
- Đa số các phương pháp Machine Learning truyền thống hoạt động tốt nhờ vào các features mà con người thiết kế phù hợp để giải quyết bài toán cụ thể.
  - Hình bên minh họa các features để giải quyết bài toán nhận dạng thực thể đối với địa danh và tên tổ chức (Finkel, 2010)

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

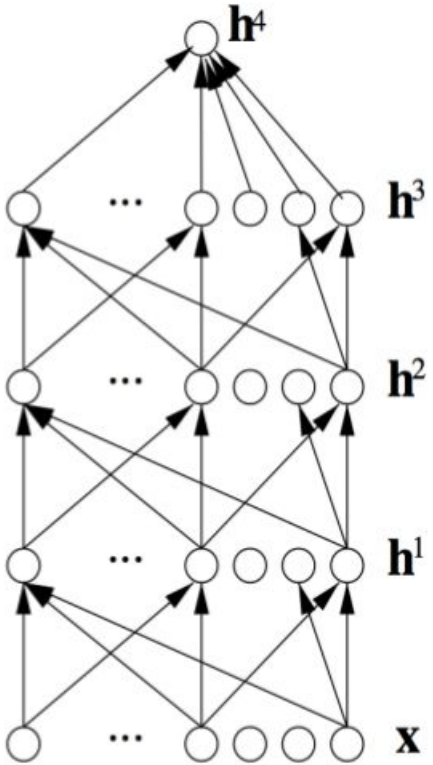
2 Deep Learning in NLP

- Deep Learning là một nhánh của Representation Learning, chú trọng vào học cách biểu diễn dữ liệu (features) tốt.
- Deep Learning sử dụng Neural Network nhiều lớp (deep) để học các tầng biểu diễn khác nhau của dữ liệu "thô" đầu vào.
- Deep Learning phát triển nhờ vào:
  - Dữ liệu lớn
  - Máy tính nhanh hơn với nhiều cores (CPUs/GPUs)
  - Mô hình, thuật toán được cải thiện



2 Deep Learning in NLP

- Các features do người thiết kế thường quá đặc trưng cho bài toán, không đầy đủ và mất nhiều thời gian để thiết kế và kiểm chứng.
- Các features được học dễ dàng thích nghi với bài toán và tự động được tìm ra một cách nhanh chóng.
- Deep Learning có thể học từ dữ liệu không gán nhãn (văn bản thô) và từ dữ liệu gán nhãn.



2 Deep Learning in NLP

- Deep NLP = Deep Learning + NLP
- Sử dụng representation learning và các phương pháp deep learning để 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:
  - Speech, Words, Syntax, Semantics
  - Part-of-speech, Named Entity Recognition, Parsing
  - Machine Translation, Sentiment Analysis, Dialogue Agents, Question Answering

2 Biểu diễn từ

Biểu diễn từ dưới dạng vector

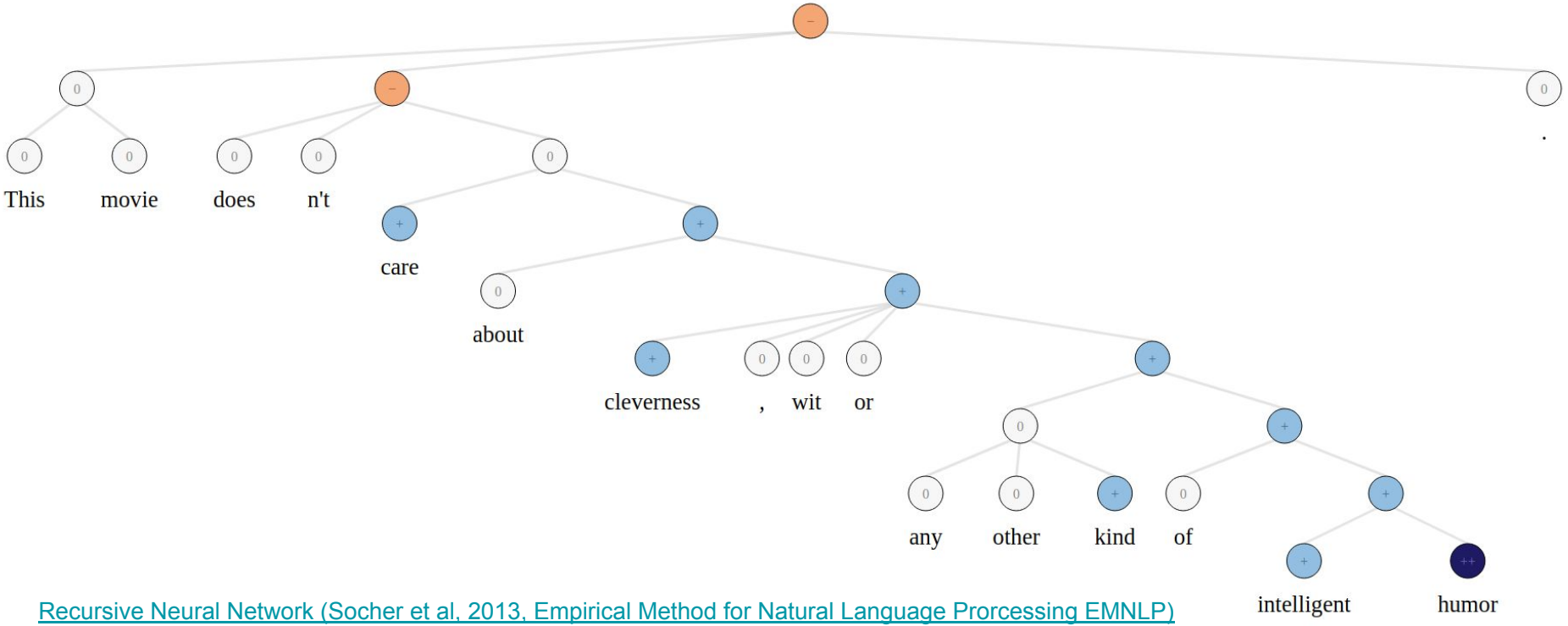
$$\text{august} = \begin{bmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.109 \\ 0.107 \\ -0.542 \\ 0.349 \end{bmatrix}$$

2 Biểu diễn từ

- Đây là top 7 từ "gần nhất" (có cosine similarity cao nhất) 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

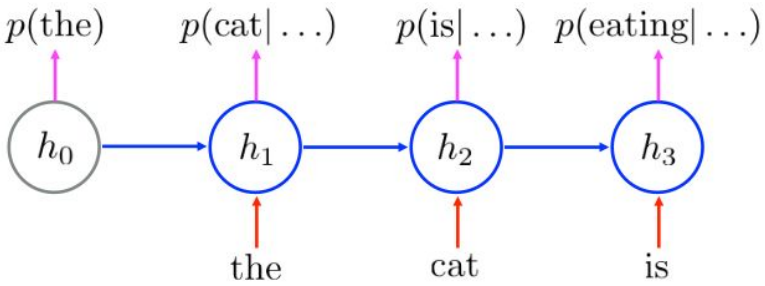
2 Sentiment Analysis



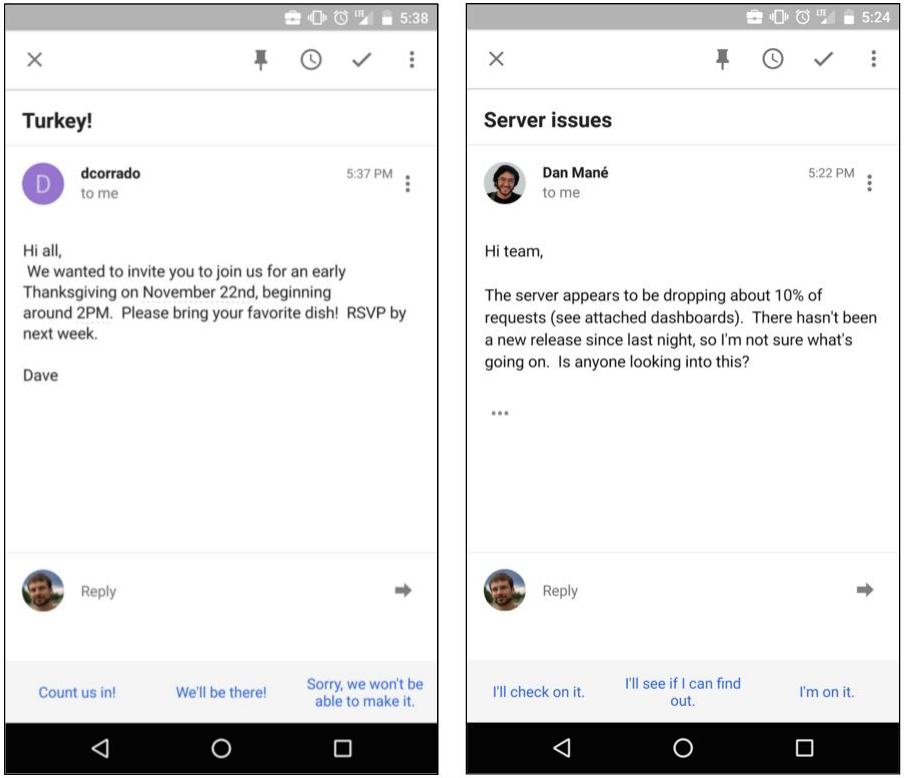
Recursive Neural Network (Socher et al, 2013, Empirical Method for Natural Language Prorcessing EMNLP)

2 Dialogue Agent và Response Generation

- Chức năng *smart reply* trên Google Inbox/GMail.

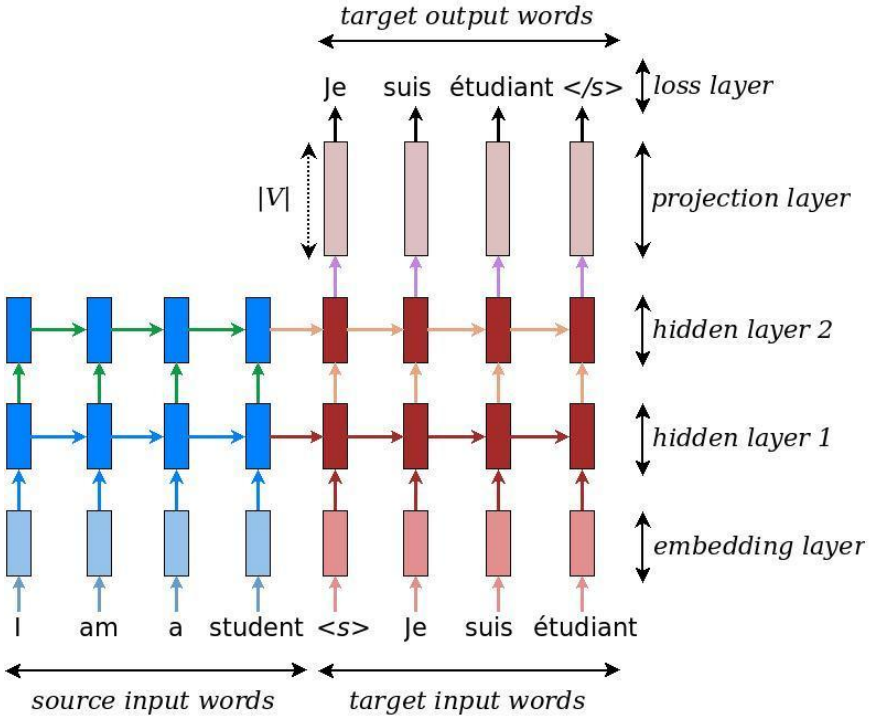


Neural Language Model



2 Machine Translation

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



Word2Vec

3 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

3 Word2Vec

- 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:
  - N-grams
  - Túi từ (Bag-of-words BOW)
  - 1-of-N, one-hot coding
  - Latent Semantic Analysis (LSA)
  - Latent Dirichlet Allocation (LDA)
  - ***Distributed Representation***

3 Word2Vec

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Biểu diễn dưới dạng "túi từ" (BOW)

motel [0.06, -0.01, 0.13, 0.07, -0.06, -0.04, 0, -0.04]

hotel [0.07, -0.03, 0.07, 0.06, -0.06, -0.03, 0.01, -0.05]

Biểu diễn dưới dạng Distributed Representation

VietAI

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

(J. R. Firth, 1957)

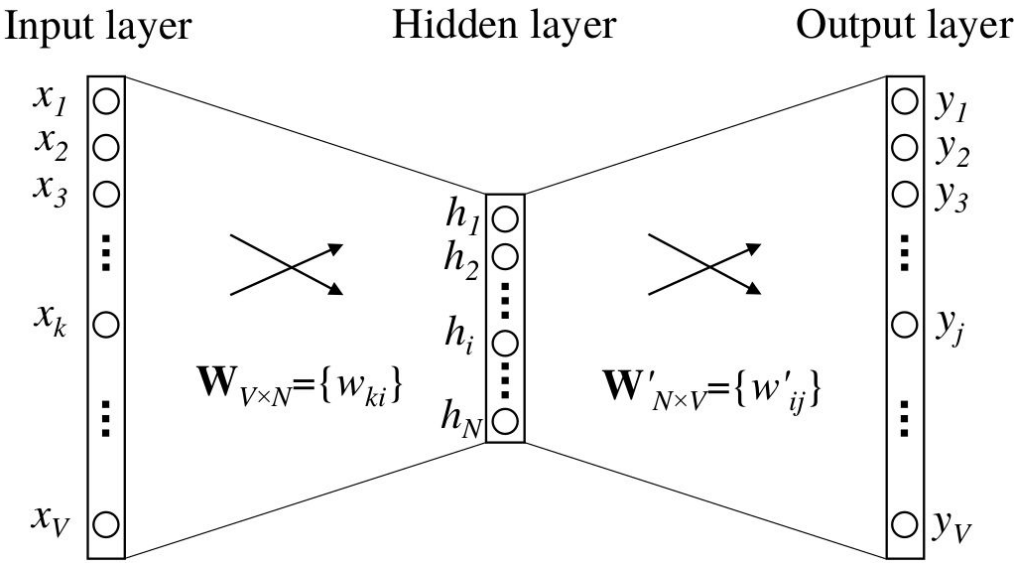
I love ***you*** so much

I love ***him*** so much

I love ***her*** so much

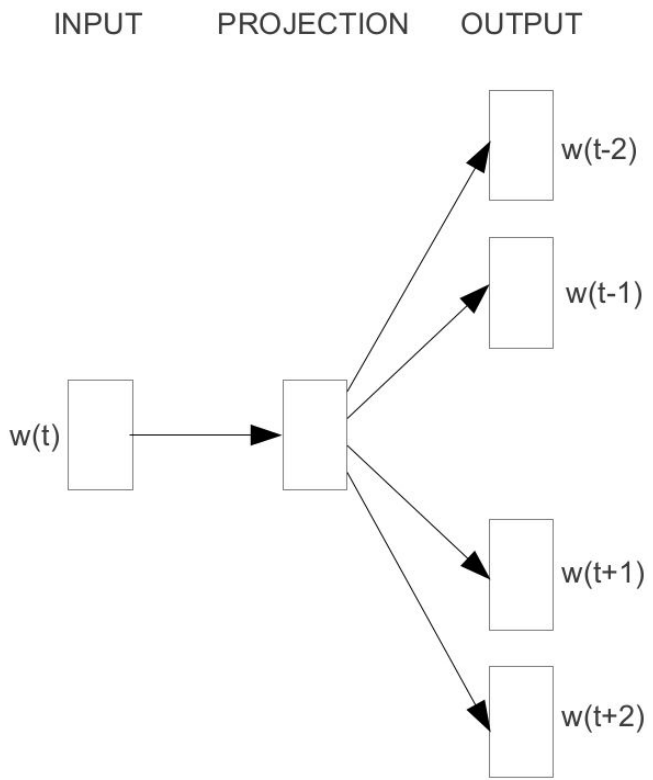
3 Giới thiệu chung về Word2Vec

- Word2Vec ra đời năm 2013 bởi Tomas Mikolov và cộng sự từ Google.
- Word2Vec bao gồm hai mô hình: Skip-gram và Continuous Bag-of-words (CBOW), cùng với đó là hai kĩ thuật để tăng tốc việc huấn luyện: Hierachical Softmax và Negative Sampling.



3 Word2Vec - Mô hình Skip-gram

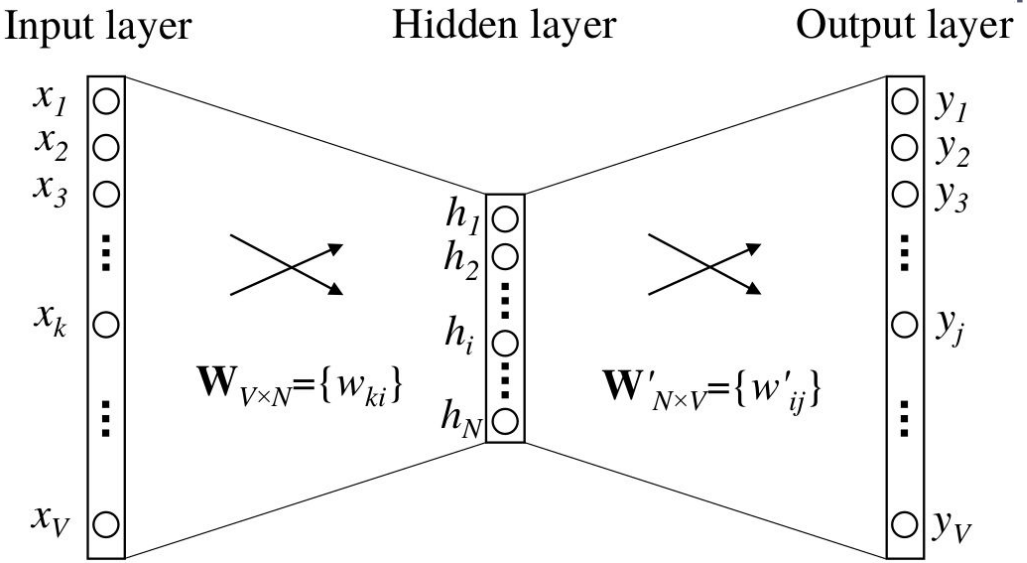
- Ý tưởng chính của Skip-gram: Dự đoán các từ xung quanh khi cho một từ ở giữa.
- Sử dụng Neural Network với 1 hidden layer và output layer là softmax.



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

3 Word2Vec - Mô hình Skip-gram

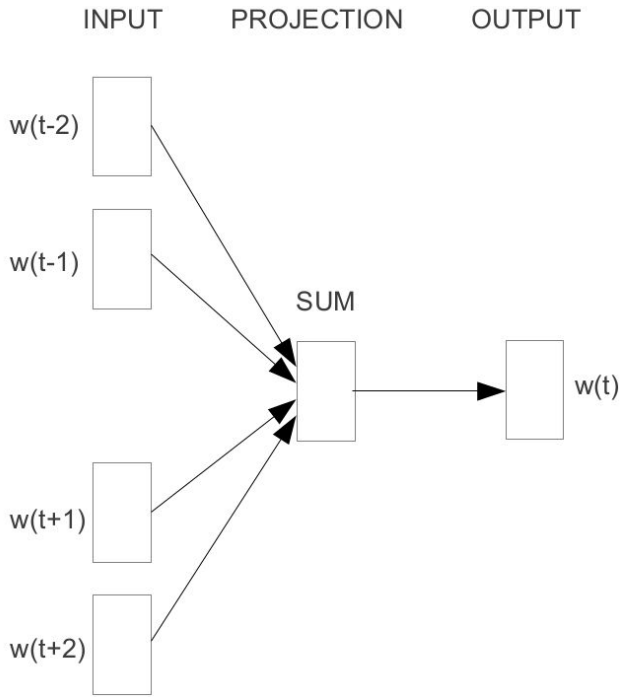
- Vector x : one-hot vector tương ứng với từ w_t
- $h = W^\top x$ việc này tương đương với lấy dòng thứ k tương ứng với từ w_t
- $y = \text{softmax}\,(h; W')$



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

3 Word2Vec - Mô hình CBOW

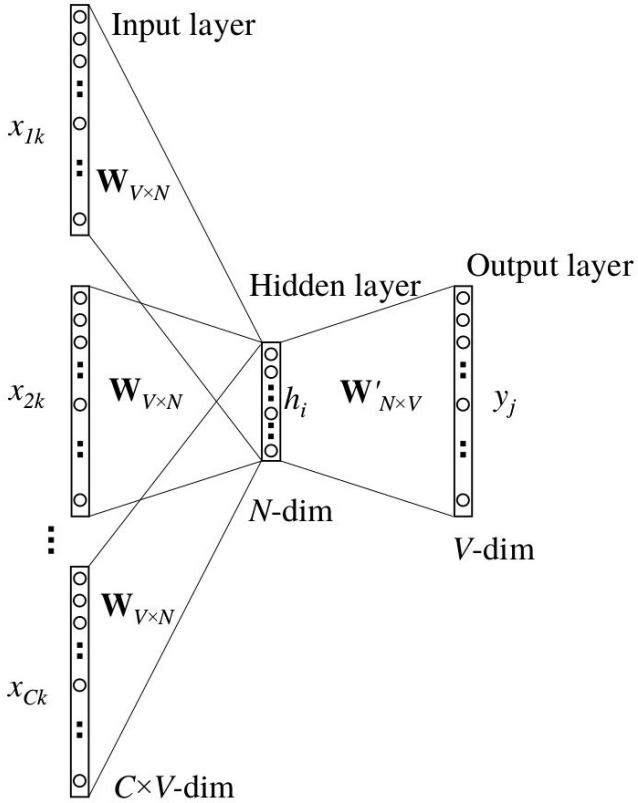
- Ý tưởng chính của Continuous Bag-of-words (CBOW): Dự đoán từ ở giữa khi cho các từ xung quanh.



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

3 Word2Vec - Mô hình CBOW

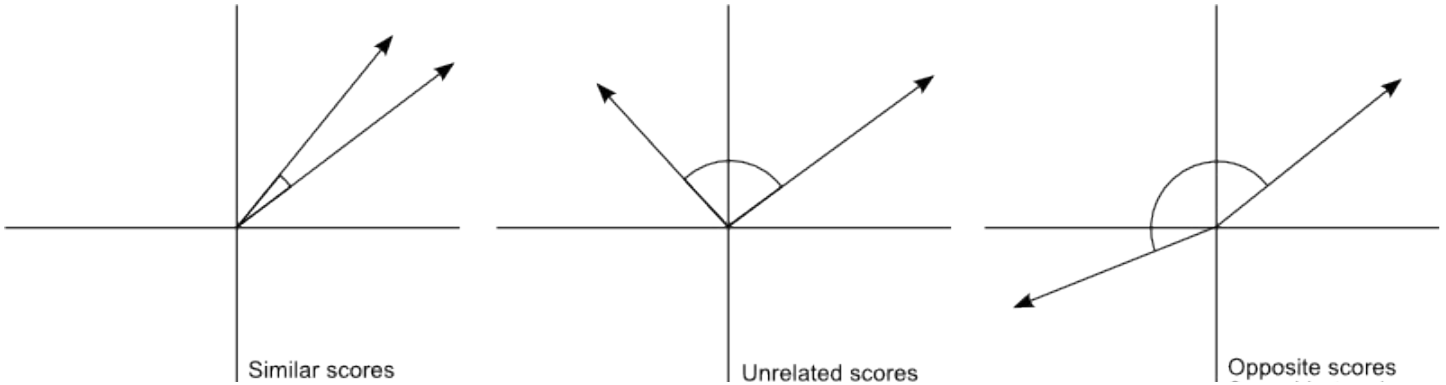
- Dựa vào các từ đầu vào tính vector h bằng cách lấy tổng hoặc trung bình của các vector tương ứng. Sau đó đưa qua softmax để dự đoán từ ở giữa.
- $h = W^\top (x_1 + x_2 + \cdots + x_C)$
- $y = \text{softmax}\,(h; W')$



3 Word2Vec - Cosine Similarity

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

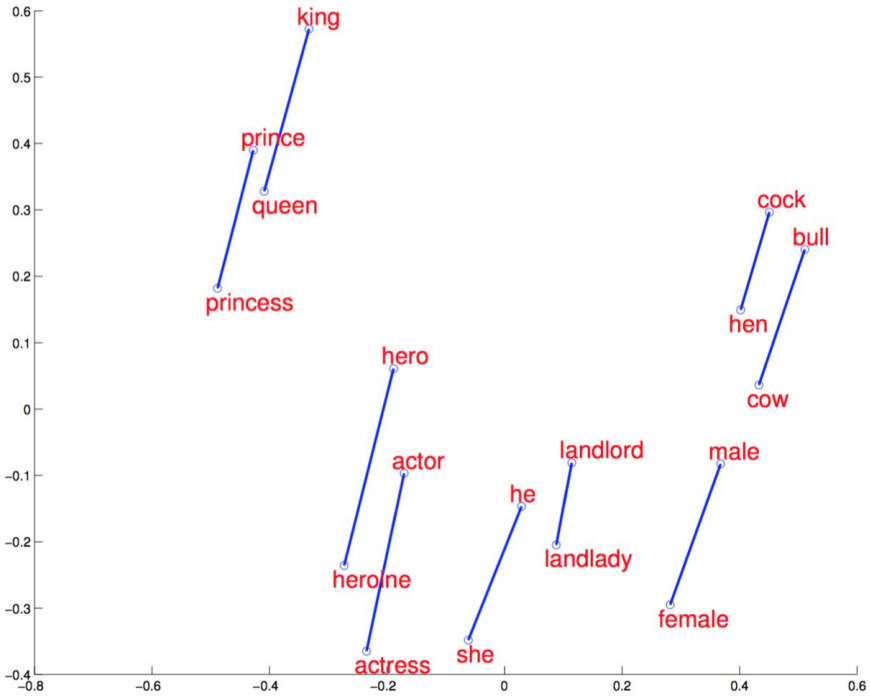
$$\text{cos-sim} = \frac{AB}{\|A\|\|B\|}$$

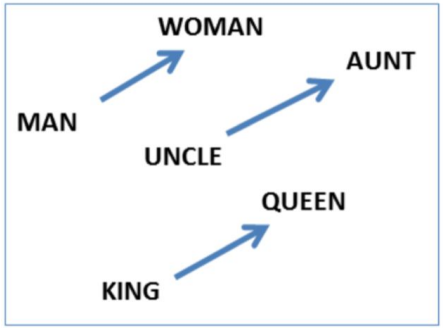


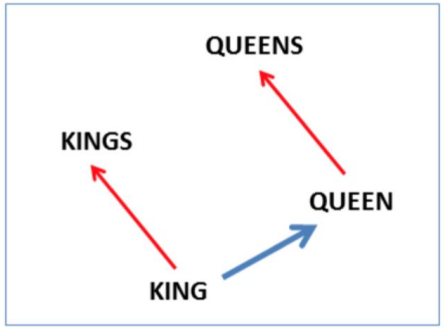
3 Word2Vec - Kết quả

- 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

3 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

3 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

3 Học vector biểu diễn các cụm từ

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

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

3 Tài liệu tham khảo

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