

# Natural Language Processing

## Introduction

### Part 1

Sudeshna Sarkar

10 Mar 2020

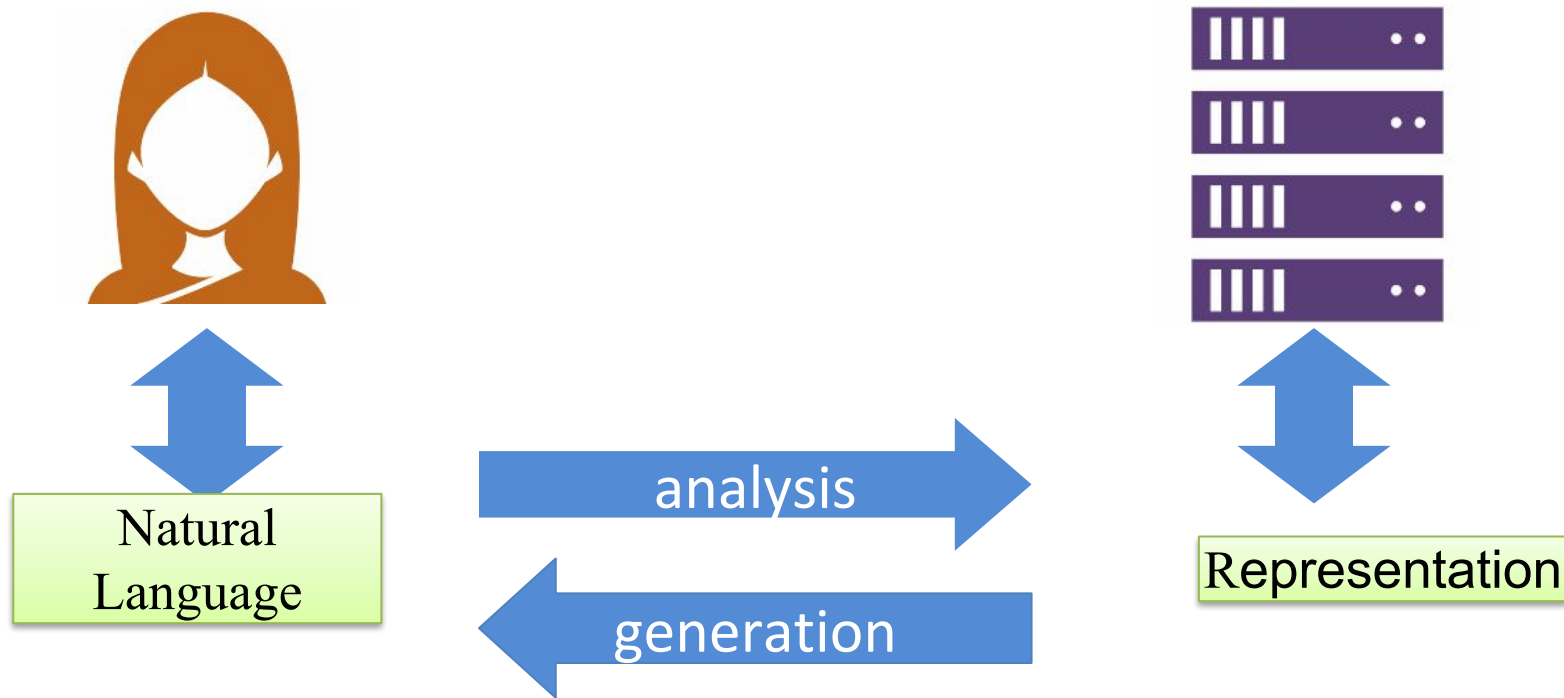
# Natural Language Processing

- NLP is focused on developing systems that allow computers to communicate with people using natural language.
- Also concerns how computational methods can aid the understanding of human language.
- Automating **Language**
  - **Analysis**      Language  $\rightarrow$  Representation
  - **Generation**    Representation  $\rightarrow$  Language
  - **Acquisition**    Obtaining the representation and necessary algorithms, from knowledge and data

# Language Processing

- Goals can be very ambitious
  - True text understanding
  - Good quality translation
- Or goals can be practical
  - Web search engines
  - Question Answering
  - Machine Translation services on the Web
  - Conversational Agents
  - Summarization
- Natural language technology not yet perfected  
But still good enough for several useful applications

# What does it mean to “know” a language?



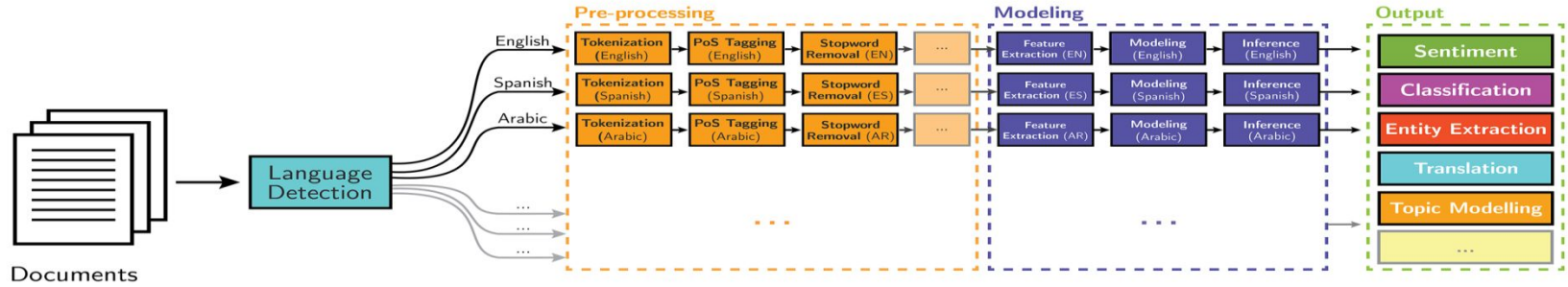
# Examples of End Systems

- Text classification
- Machine translation, information extraction, dialog interfaces, question answering...
- human-level comprehension

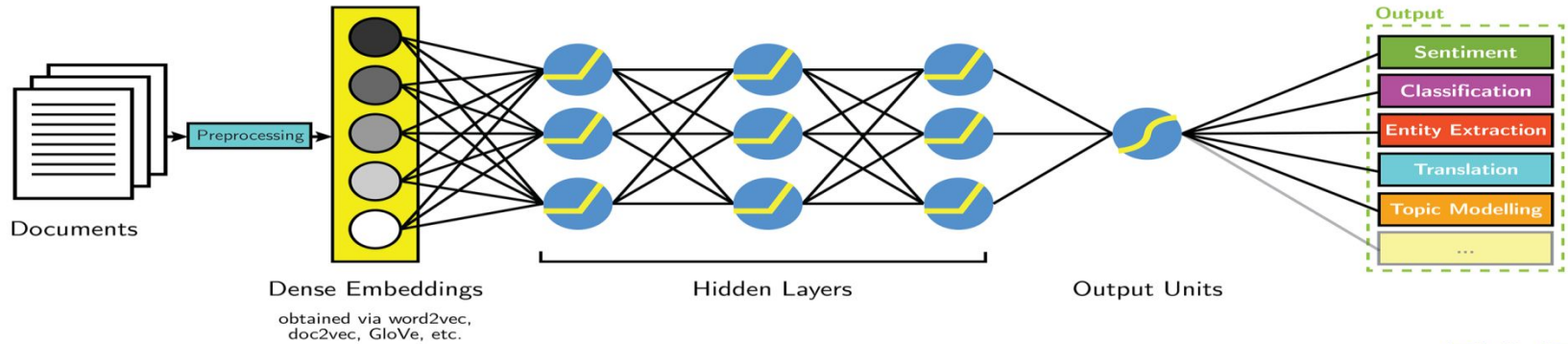
# Three Generations and Three Views

- **Hand-crafted Systems**  
–Knowledge Engineering [1950s–]
  - **Automatic, Trainable (Machine Learning)**  
Systems with engineered features [1985s–2012]
  - **Automatic, Trainable Neural architectures** with no/limited engineered features [2012--]
1. Classical View: Layered Processing; Various Ambiguities
  2. Statistical/Machine Learning View
  3. Deep Learning View

## Classical NLP



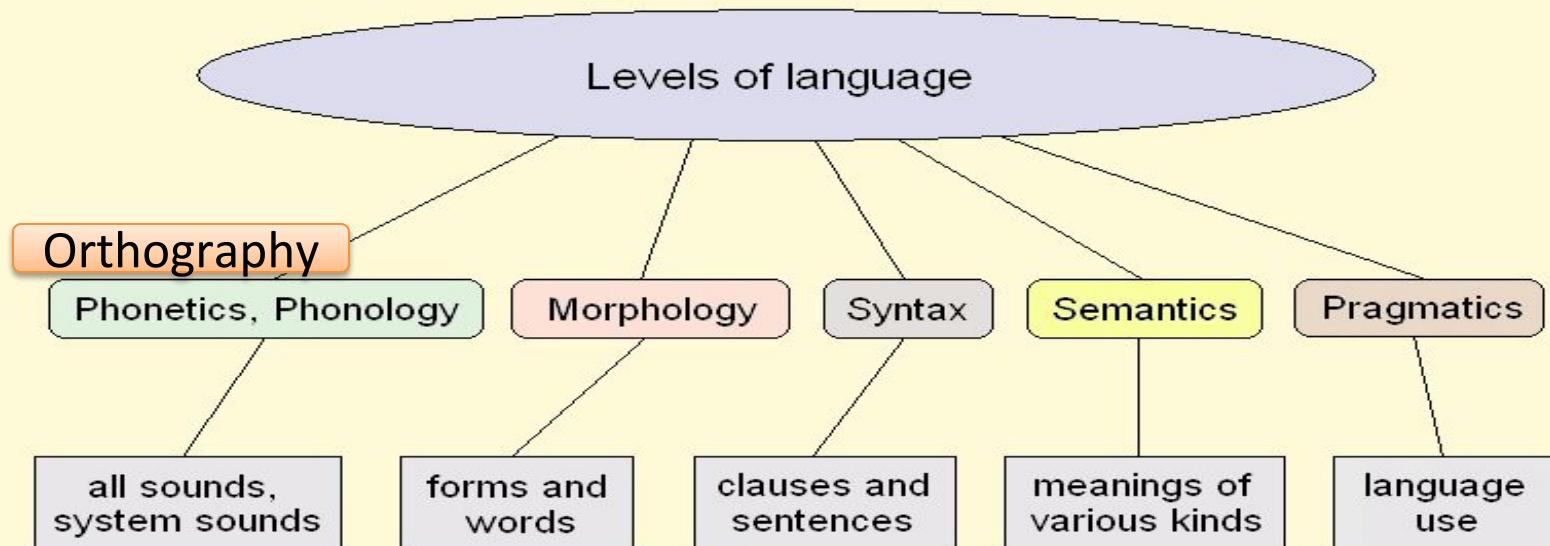
## Deep Learning-based NLP



# NLP is Hard

- Ambiguity
- Ill-defined problems
- AI-complete





# Complexity of Linguistic Representations

- Richness: there are many ways to express the same meaning, and immeasurably many meanings to express.
- Each level interacts with the others.
- There is tremendous diversity in human languages.
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

# Natural language understanding

- Uncovering the mappings between the linear sequence of words and the meaning that it encodes.
- Representing this meaning in a useful (usually symbolic) representation.
- By definition - heavily dependent on the target task
  - Words and structures mean different things in different contexts
  - The required target representation is different for different tasks.
- Appropriateness of a representation depends on the application.

# Why is NLP Hard?

- The mappings between words, their linguistic structure and the meaning that they encode is extremely complex and difficult to model and decompose.
- Natural language is very ambiguous
  - **Lexical (word level) ambiguity** -- different meanings of words
  - **Syntactic ambiguity** -- different ways to parse the sentence
  - **Interpreting partial information** -- how to interpret pronouns
  - **Contextual information** -- context of the sentence may affect the meaning of that sentence.
- Noisy Input

# Complexity of Linguistic Representations

- Richness: there are many ways to express the same meaning, and immeasurably many meanings to express.
- Each level interacts with the others.
- There is tremendous diversity in human languages.
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

# Components of NLP

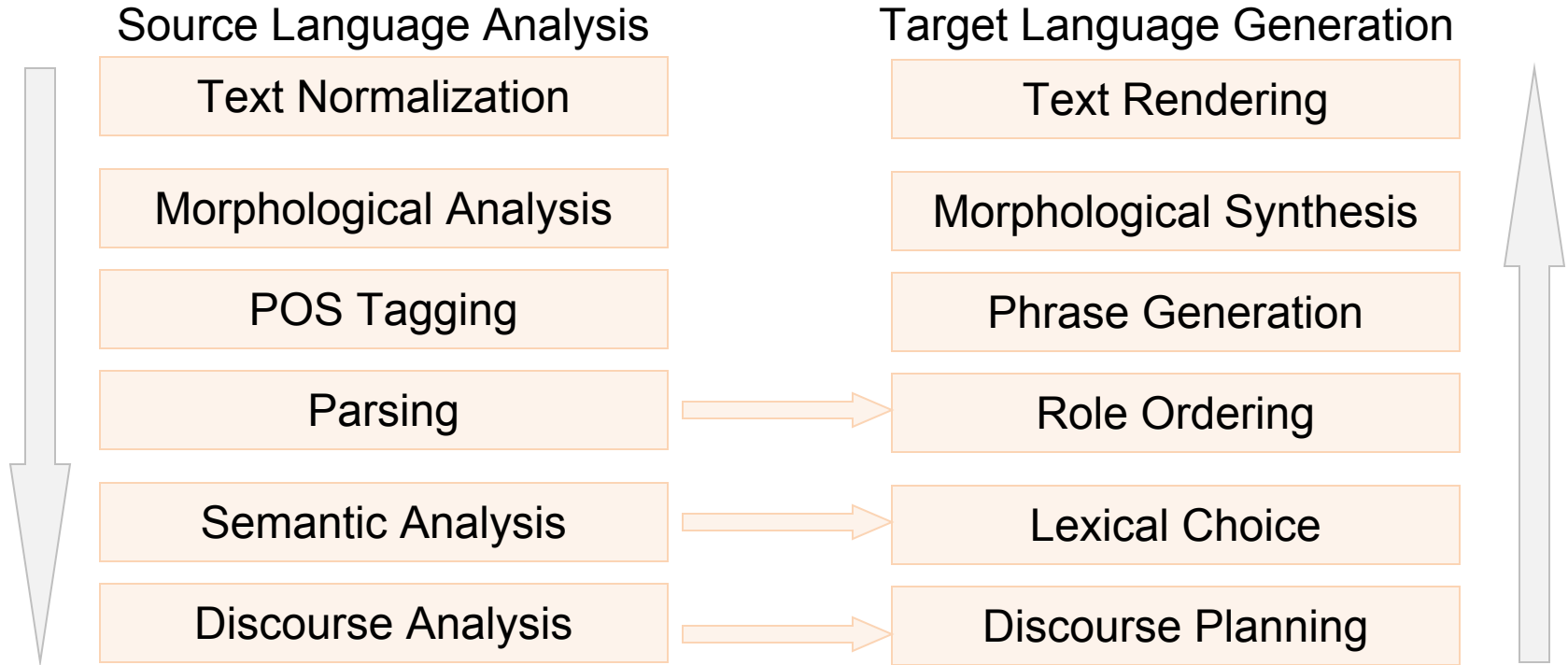
- **Natural Language Understanding**

- Mapping the given input in the natural language into a useful representation.
- Different level of analysis required: morphological, syntactic, semantic, discourse

- **Natural Language Generation**

- Producing output in the natural language from some internal representation.
- Different level of synthesis required:
  - deep planning (what to say),
  - syntactic generation

# The Reductionist Approach



# Morphology

- The identification, analysis and description of the structure of words
- Challenges
  - Ambiguity (flies, bears, মাতাল)
  - Segmenting text into words (Thai)
  - Sandhi splitting (Sanskrit)
  - Morphological variations
  - Words with multiple meanings (based on context, domain)
  - Multiword expression



# Syntax

- Syntax concerns the way in which words can be combined together to form (grammatical) sentences
  1. revolutionary new ideas appear infrequently
  2. colourless green ideas sleep furiously
  3. \*ideas green furiously colourless sleep
- Words combine syntactically in certain orders in a way which mirrors the meaning conveyed
- John gave her dog biscuits
  - (john (gave (her) (dog biscuits)))
  - (john (gave (her dog) (biscuits)))

# Semantics

- The manner in which lexical meaning is combined morphologically and syntactically to form the meaning of a sentence
  - Concerns the meaning of words, phrases and sentences
  - The meaning of a sentence is usually a productive combination of the meaning of its words

# Discourse analysis

- The meaning of a sentence depends upon the sentences that preceded it and also invokes the meaning of the sentences that follow it.
- The discourse structure of connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast)

# Hardness of NLP

- Mappings across levels are complex.
  - A string may have many possible interpretations in different contexts, and resolving ambiguity correctly may rely on knowing a lot about the world.
  - Richness: any meaning may be expressed many ways, and there are immeasurably many meanings.
  - Linguistic diversity across languages, dialects, genres, styles

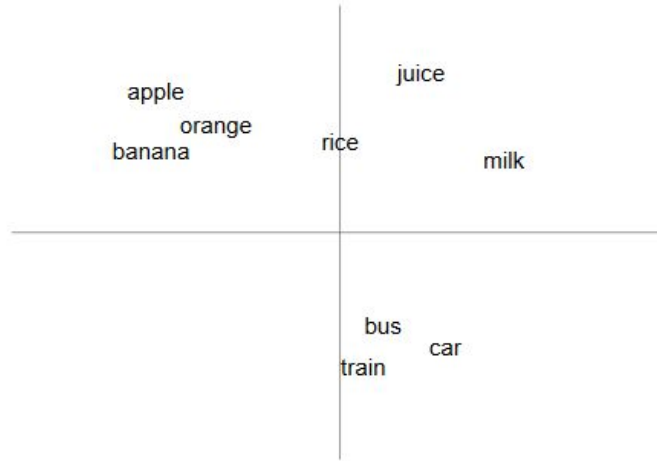
# Natural Language Processing

## Word Representation

Sudeshna Sarkar

10 Mar 2020

***“A word is known by the company it keeps”***



# Word Representation

- Continuous Representation: based on context
- Distributional hypothesis

You can get a lot of value by representing a word by means of its neighbors

“You shall know a word by the company it keeps”

(J. R. Firth 1957: 11)

One of the most successful ideas of modern NLP

government debt problems turning into banking crises as has happened in  
saying that Europe needs unified banking regulation to replace the hodgepodge

↩ These words will represent *banking* ↗

# Representation

- We need effective representation of :
  - Words, Sentences, Text

1: Use existing thesauri or ontologies like WordNet

Drawbacks:

- Manual
- Not context specific

2: Use co-occurrences for word similarity. Drawbacks:

- Quadratic space needed
- Relative position and order of words not considered



## word2vec approach to represent the meaning of word

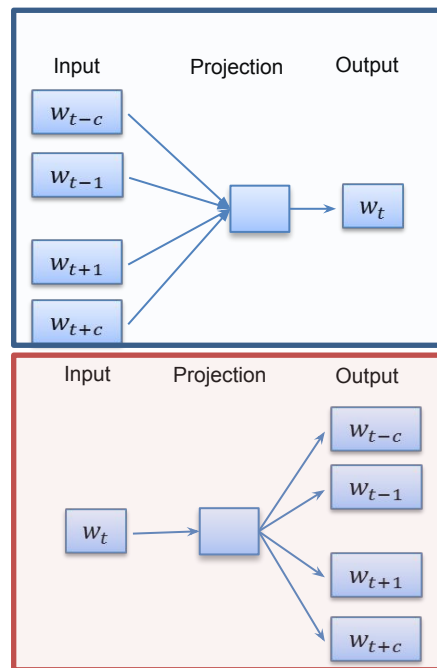
- Represent each word with a low-dimensional vector
- Word similarity = vector similarity
- Key idea: Predict surrounding words of every word
- Faster and can easily incorporate a new sentence/document or add a word to the vocabulary

# Word2vec

- Representation of words

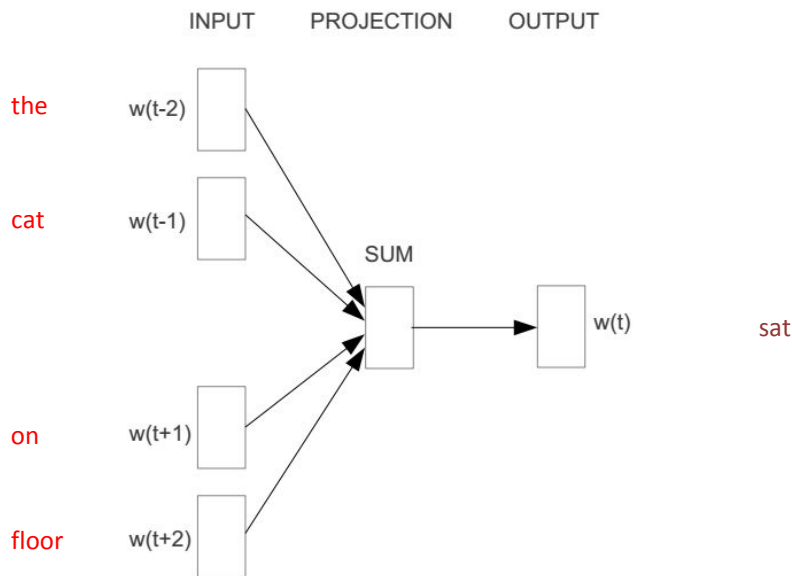
“Similar words have similar contexts”

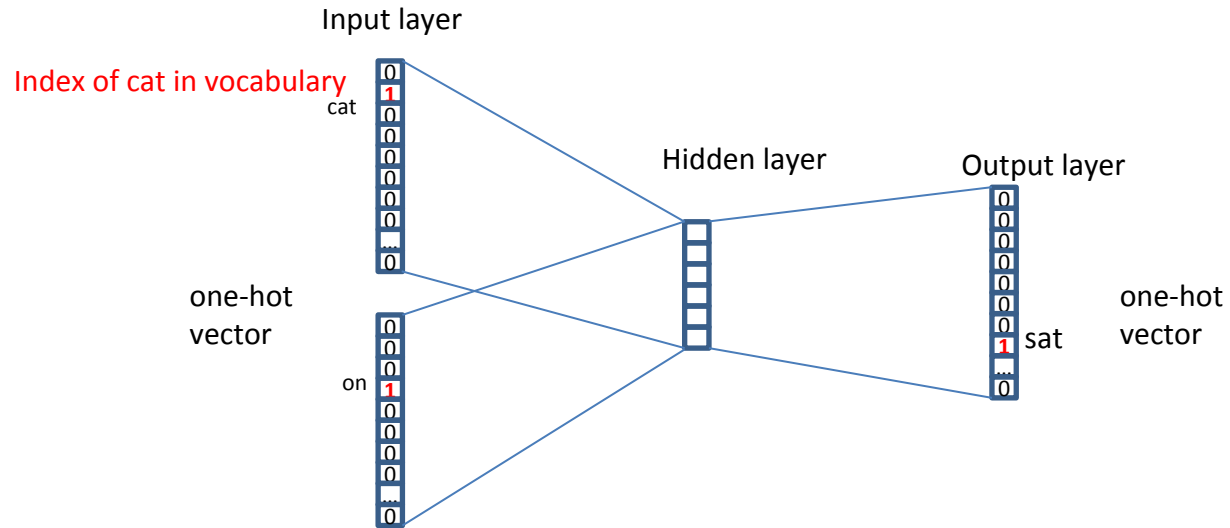
1. CBOW:  $P(\text{Word}|\text{Context})$
2. Skipgram:  $P(\text{Context}|\text{Word})$

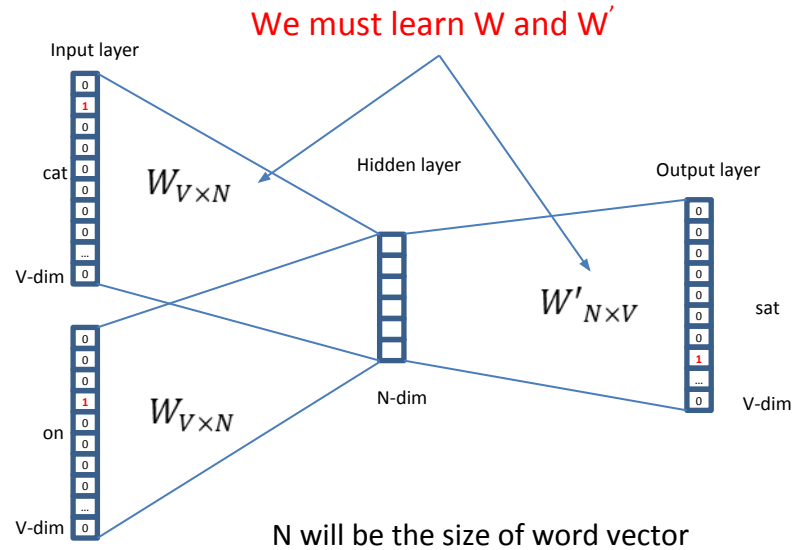


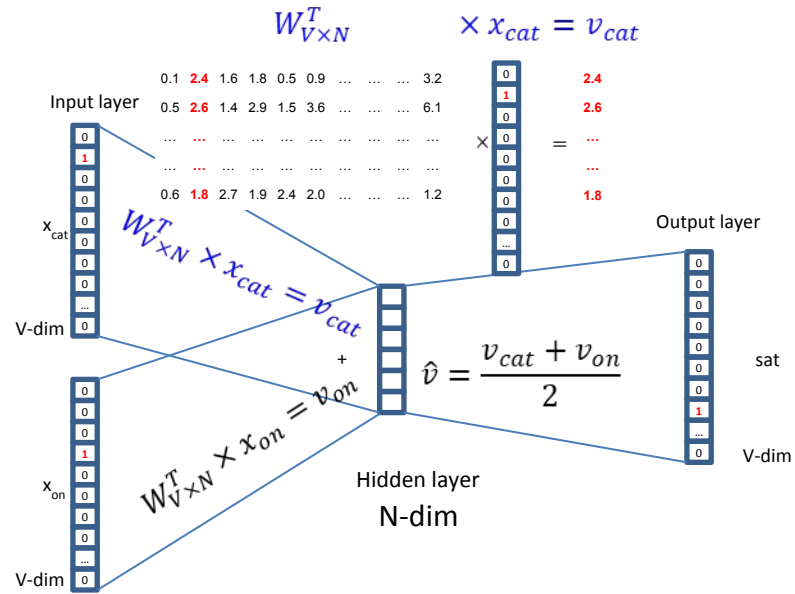
# Word2vec – Continuous Bag of Word

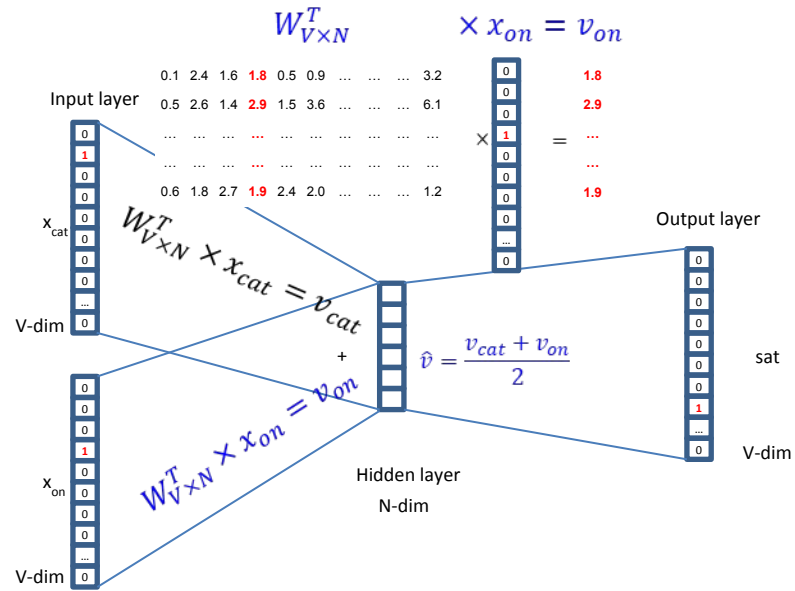
- E.g. “The cat sat on floor”
  - Window size = 2

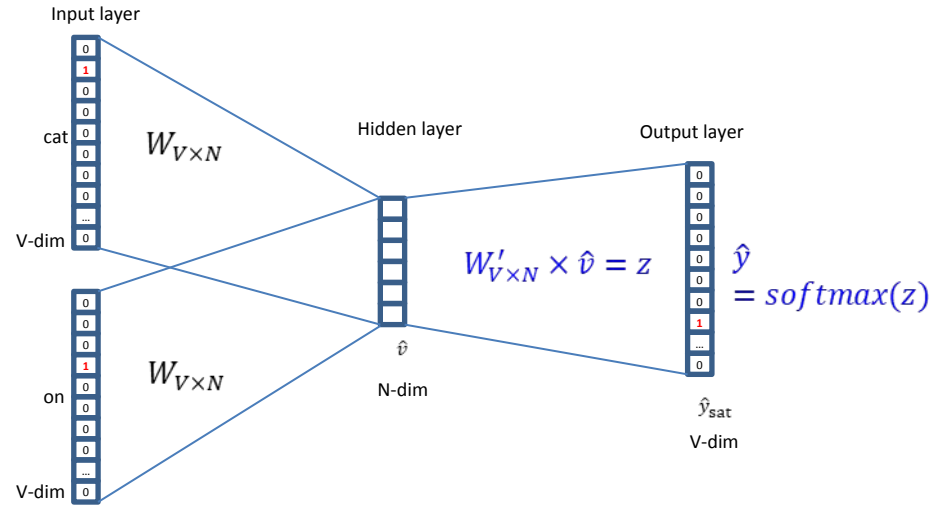




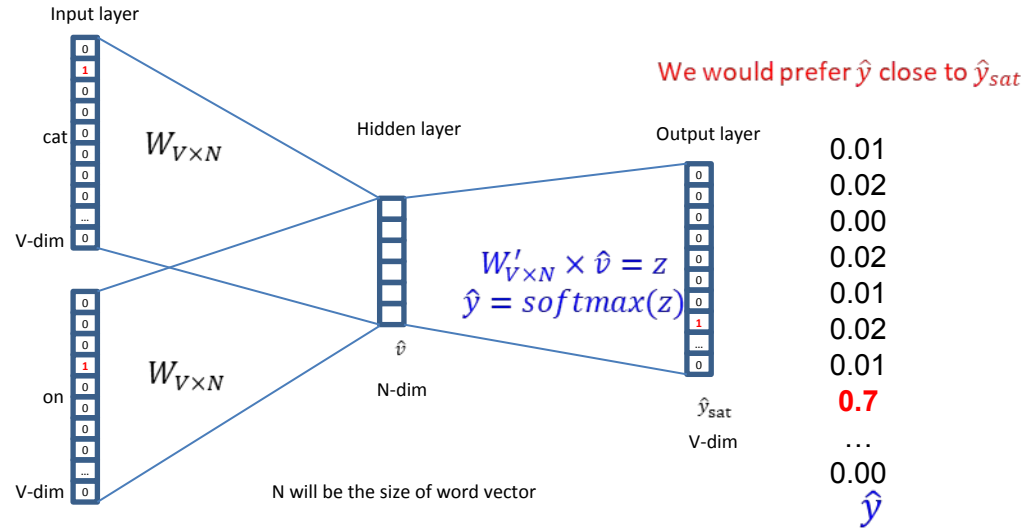


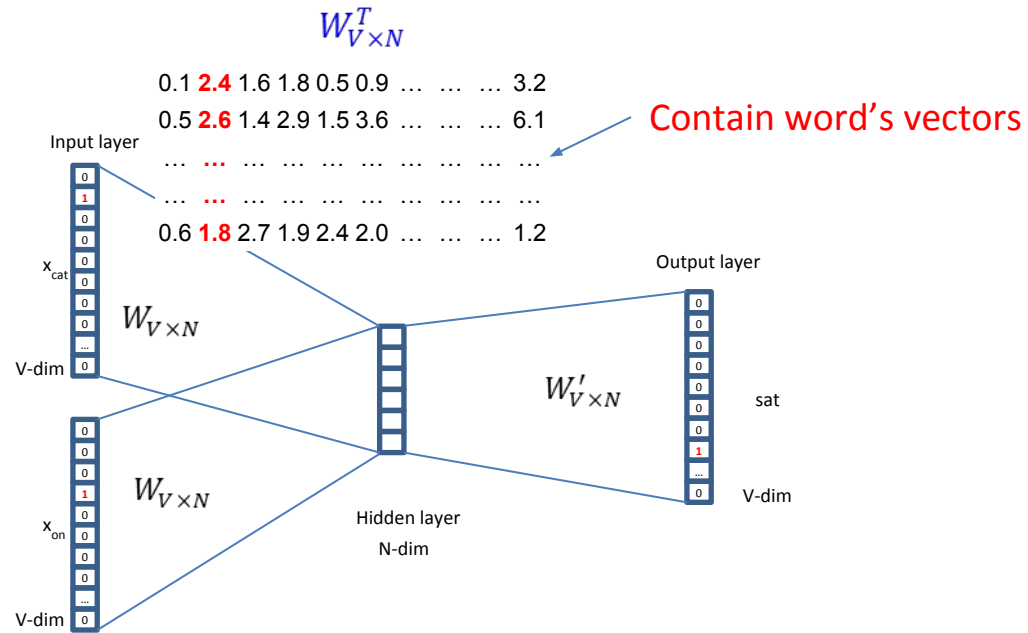












We can consider either  $W$  or  $W'$  as the word's representation. Or even take the average.

# Some interesting results

## Word Analogies

Test for linear relationships, examined by Mikolov et al. (2014)

a:b :: c:?



$$d = \arg \max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

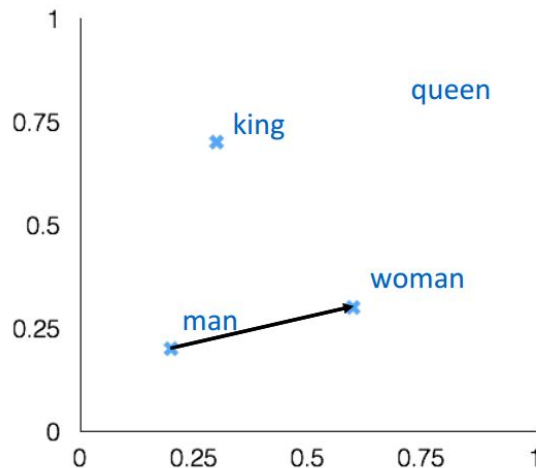
+ king [ 0.30 0.70 ]

- man [ 0.20 0.20 ]

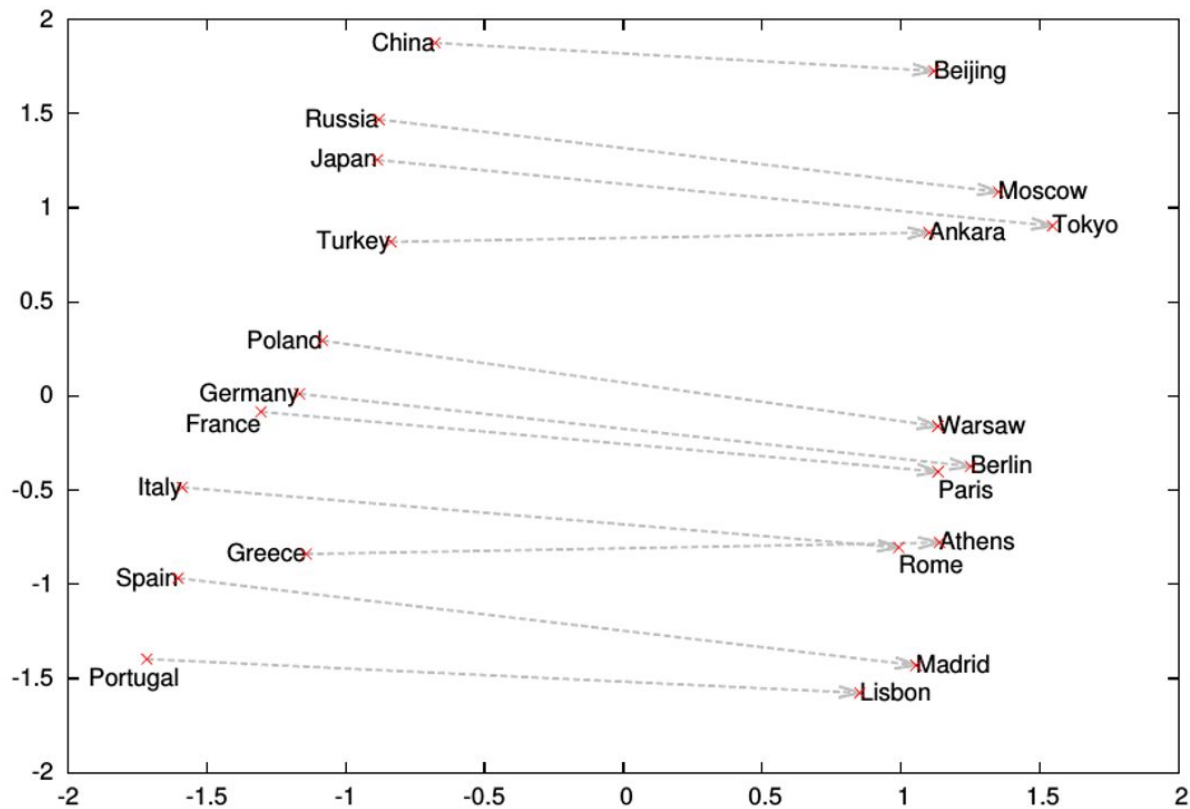
+ woman [ 0.60 0.30 ]

---

queen [ 0.70 0.80 ]



# Word analogies



# Word2Vec Objective

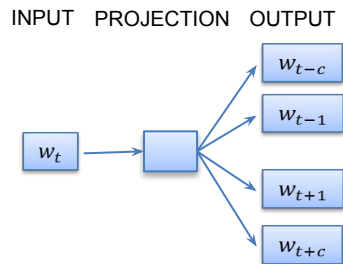
$$J(\theta) = \frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} \log p(w_j | c)$$

# Word2Vec Objective

$$p(w_j|c) = \frac{\exp(w_j^T c)}{\sum_{i=1}^N \exp(w_i^T c)}$$

# Skipgram Model

- Input: Central word  $w_t$
- Output: Words in its context:  $w_{\text{con}}$   
 $\{w_{t-c}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+c}\}$
- Each input word represented by a 1-hot encoding of size V



Source Text:

**Deep Learning attempts to learn multiple levels of representation from data.**

Input output pairs :

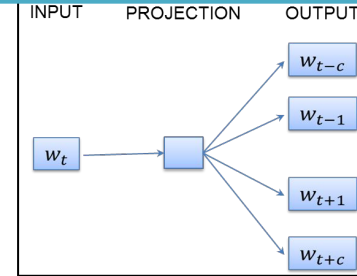
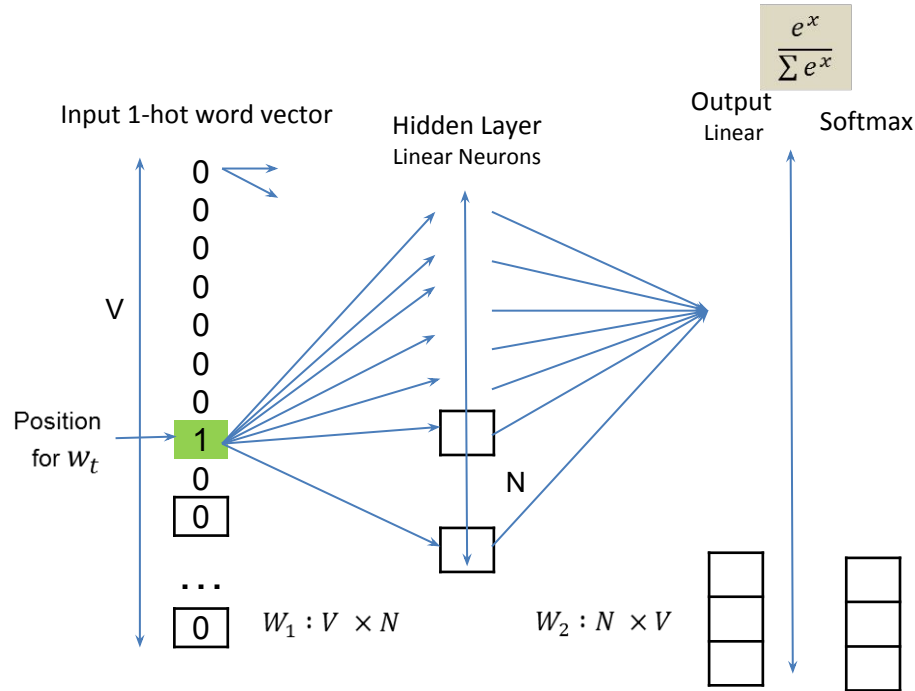
**Positive samples:**

(representation, levels)  
(representation, of)  
(representation, from)  
(representation, data)

**Negative samples:**

(representation, x)  
[x: all other words except the 4 positive]

# Skipgram Model



Probability that the word in a context position is  $w_i$

Positive sampling  
Negative sampling



# Skipgram: Loss function

- Maximize 
$$\frac{1}{T} \sum_{t=1}^T \sum_{\text{context}} \log p(w_{\text{context}} | w_t)$$

$p(w_{\text{con}} | w_t)$  is the output of softmax classifier

$$p(w_{\text{con}} | w_t) = \frac{\exp(v'_{w_{\text{con}}} \cdot v_{w_t})}{\sum_{w=1}^W \exp(v'_w \cdot v_{w_t})}$$

Let the model parameters be  $\theta$ . The solution is given by

$$\begin{aligned} & \underset{\theta}{\operatorname{argmax}} \sum_{(w_t, w_c) \in D} \log p(w_{\text{con}} | w_t; \theta) \\ &= \sum_{(w_t, w_c) \in D} \left( \log e^{(v'_{w_{\text{con}}} \cdot v_{w_t})} - \log \sum_x e^{(v'_x \cdot v_{w_t})} \right) \end{aligned}$$

**Time  $O(V)$**

V: vocabulary size

**Improve Efficiency**

**1. Hierarchical softmax:**

$O(\log V)$

# Skipgram: Loss function

● Maximize  $\frac{1}{T} \sum_{t=1}^T \sum_{\text{context}} \log p(w_{\text{context}} | w_t)$

$p(w_{\text{con}} | w_t)$  is the output of softmax classifier

$$p(w_{\text{con}} | w_t) = \frac{\exp(v'_{w_{\text{con}}} \cdot v_{w_t})}{\sum_{w=1}^W \exp(v'_w \cdot v_{w_t})}$$

Let the model parameters be  $\theta$ . The solution is given by

$$\begin{aligned} & \underset{\theta}{\operatorname{argmax}} \sum_{(w_t, w_c) \in D} \log p(w_{\text{con}} | w_t; \theta) \\ &= \sum_{(w_t, w_c) \in D} \left( \log e^{(v'_{w_{\text{con}}} \cdot v_{w_t})} - \log \sum_x e^{(v'_x \cdot v_{w_t})} \right) \end{aligned}$$

**2. Negative sampling:** Sample instead of taking all contexts into account

$$\sum_{(w_t, w_c) \in D} \left( \log \sigma(v'_{w_{\text{con}}} \cdot v_{w_t}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} [\log \sigma(-v'_{w_i} \cdot v_{w_t})] \right)$$

Subsampling of frequent words

# Skip-Grams with Negative Sampling (SGNS)

Marco saw a furry little **wampimuk** hiding in the tree.

# Skip-Grams with Negative Sampling (SGNS)

Marco saw a furry little wampimuk hiding in the tree.

## words

wampimuk

wampimuk

wampimuk

wampimuk

...

## contexts


furry

little

hiding

in

...



$D$  (data)

“word2vec  
Explained...”  
Goldberg & Levy, arXiv  
2014

# Skip-Grams with Negative Sampling (SGNS)

♣ **Maximize:**  $\sigma(\vec{w} \cdot \vec{c})$

- $c$  was **observed** with  $w$

words

wampimuk

wampimuk

wampimuk

wampimuk

contexts

furry

little

hiding

in

♣ **Minimize:**  $\sigma(\vec{w} \cdot \vec{c}')$

- $c'$  was **hallucinated** with  $w$

words

wampimuk

wampimuk

wampimuk

wampimuk

contexts

Australia

cyber

the

1985