Introduction to Text Representations

LING575

Yan Song

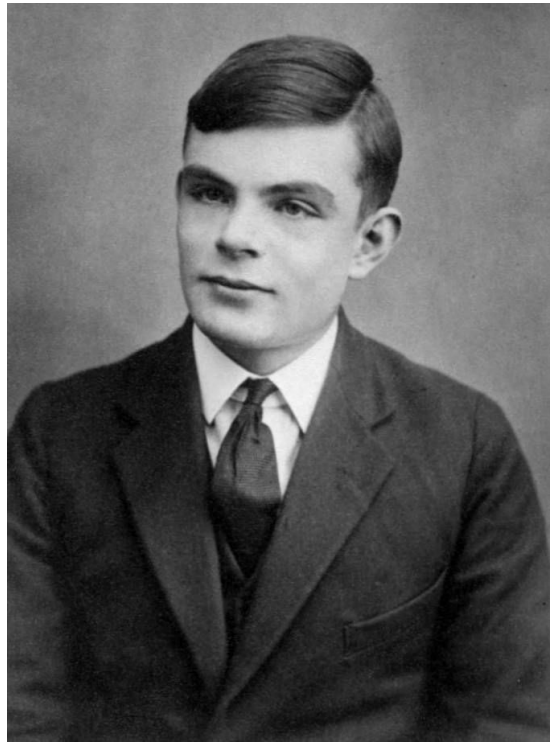
Outline

- Why?
- Encodings for Computer
- Different Represeantions for Natural Languages
  - One-hot Representation
  - Other Numerical Ones
  - Neural Representation
    - Evaluate
- Summary

Why?

- “Can machines think?”

  Computing Machinery and Intelligence, 1950
- “replace the question by another, which is closely related to it and is expressed in relatively unambiguous words”
  - Define “think”
  - Define “machine”, with the tools



Alan Turing

Why?

- Natural Language Processing
  - 1950-1966 ALPAC report
  - 1966-1990 Rationalism
  - 1990-2010 Empericalism
  - 2010-nowadays Learning and Data Domination
- Methodology
  - Rule-based
  - Example-based
  - Statistical methods
    - Shallow learning
    - Deep learning

Why?

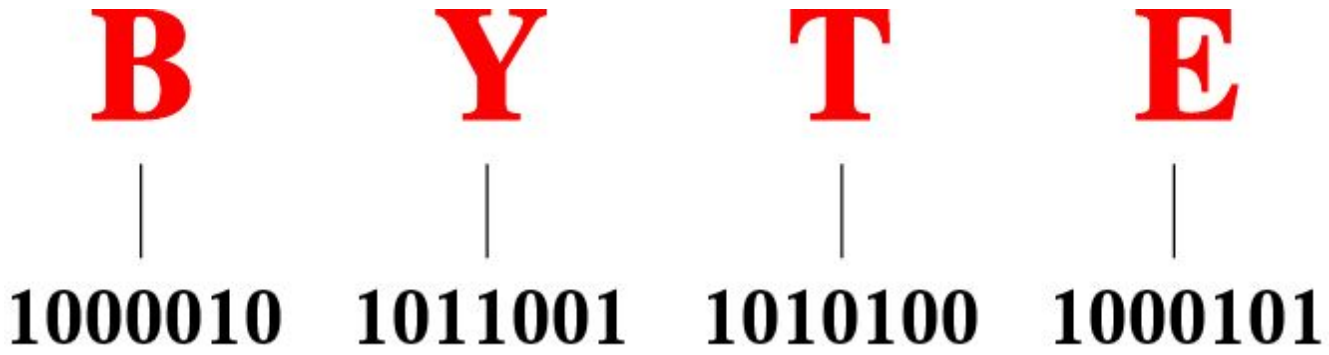
- (Written) Language is a symbol system
- Machine normally deals with numbers (0,1)
- Keep as much information as possible
- Language independent
- Efficiency and robustness
- ...

Encodings for Computer

- What do you have in your mind?

Encodings for Computer

- What computer systems do in coding texts?



(1) ASCII

- In ASCII codes each code is made of 7 bits.
- Number of possible codes M = 2^7 = 128 codes.
- Bit-patterns ranging from 0000000 to 1111111
- The first pattern represents (null character)
- The last pattern represents (delete character)

(2) Extended ASCII

- Is invented to make the bit-pattern length equal to 8 bits (Byte), by adding a bit to the left of the ASCII code representation.

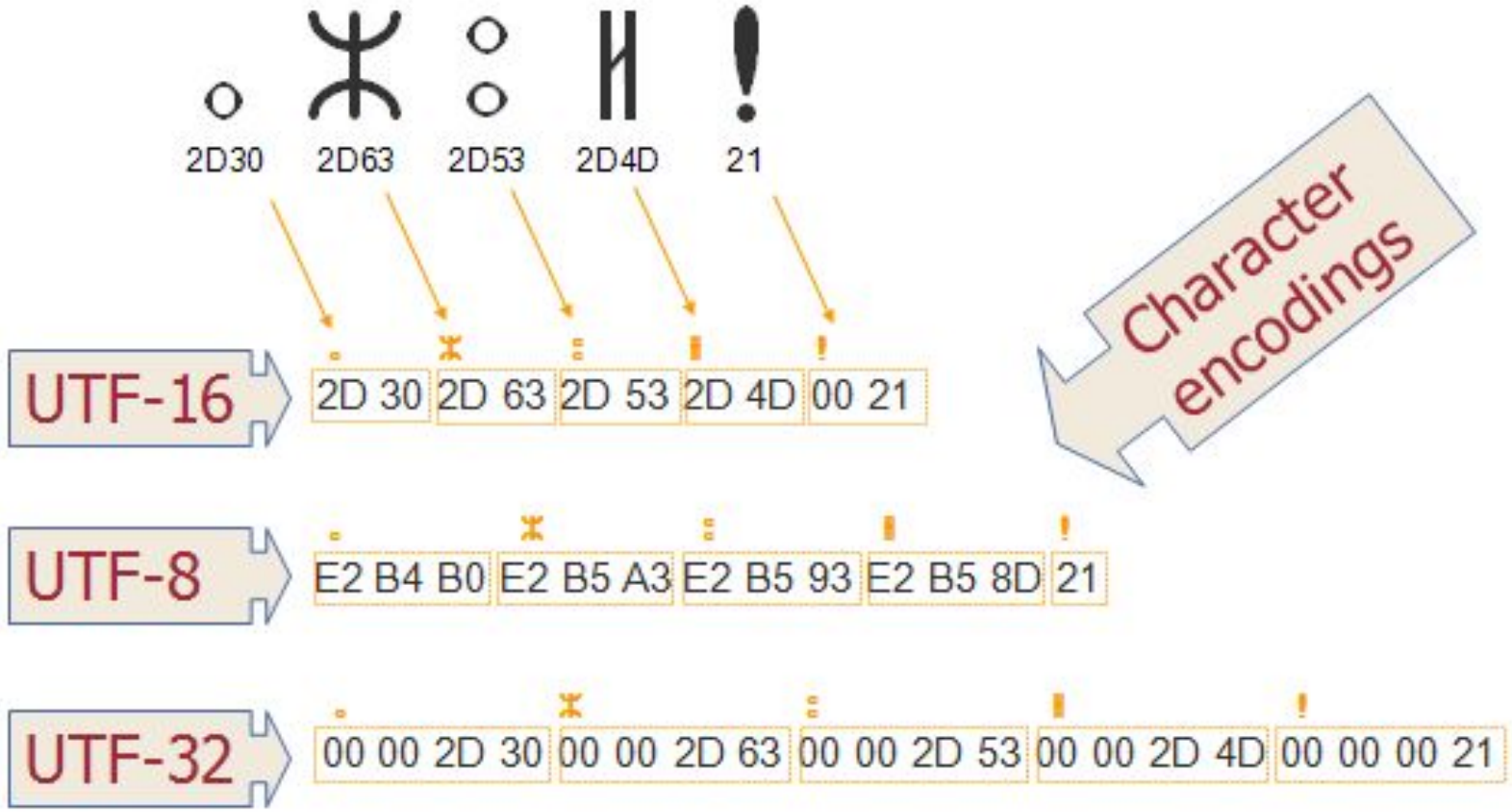
  Ex. If ASCII code is 1111111 the extended ASCII code is 01111111.
- Extended ASCII is not used because it is not standardized as each manufacturer has different 8-bits system.

(3) Unicode

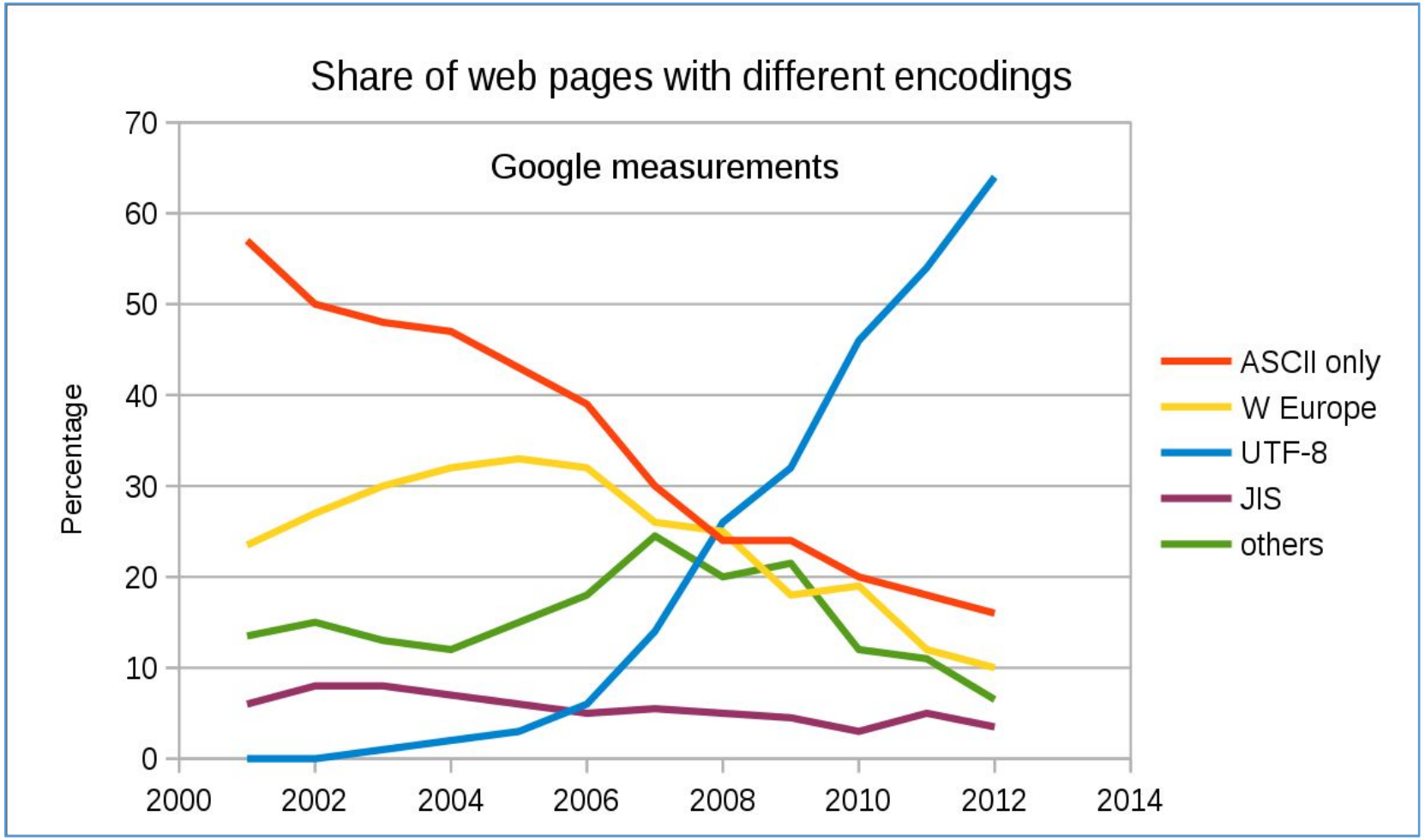
- To represent more languages' character beside English, Unicode is invented
- Uses 16 bit pattern $\rightarrow$ # of codes = 2^{16}=65536

  (almost) enough to represent all world's languages.
- Some codes are allocated for geographical and special symbols
- Java uses Unicode, Microsoft uses the first 256 symbols

(3) Unicode - UTF



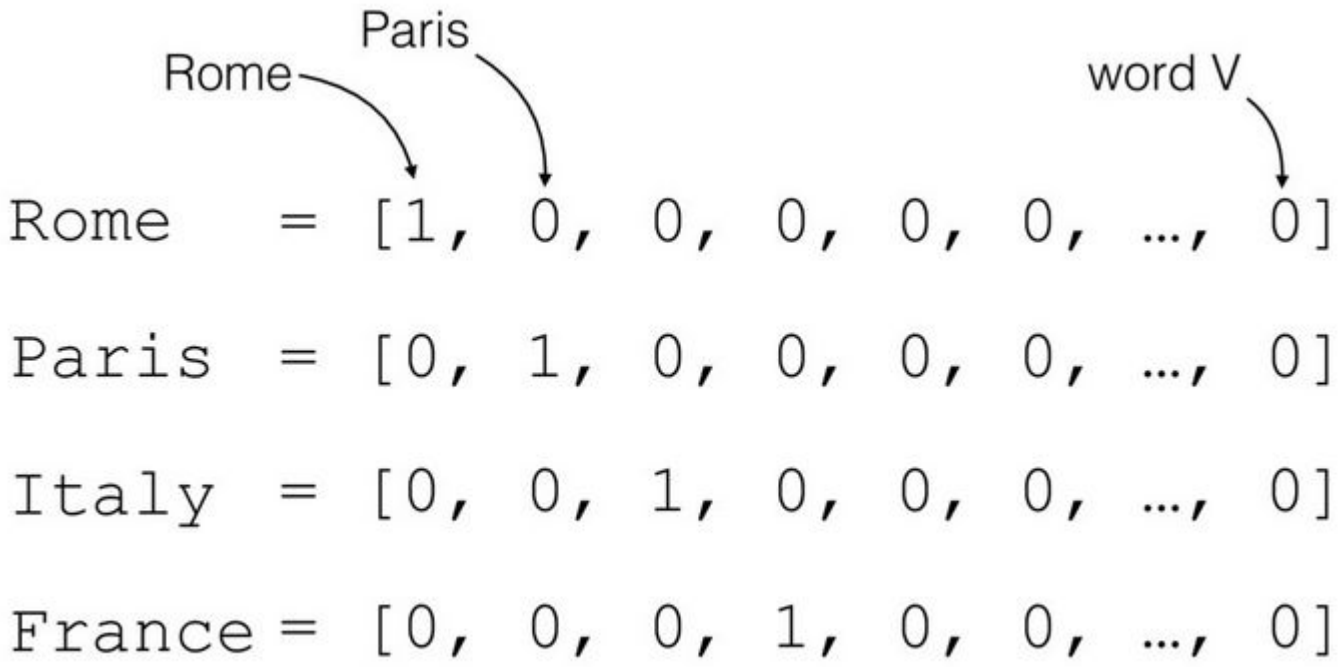
(3) Unicode - UTF-8



What we need to present natural language?

- Can we use the aforementioned coding schemes in NLP
  - Yes?
  - No?
- What can be taken away?
  - One array or sequence
  - Numerical
  - Discriminative (discreet)

One-hot Representation



One-hot Representation

- Advantages?
  - Intuitionistic
  - Easy to encode
  - Categorical meaning
  - ...

One-hot Representation

- Disadvantages?
  - Serious data sparseness
  - No word-word relation
  - Restricted in further utilization
  - ...

More Numerical?

- What if the “1” is an “*N*”?
- What if there are N “1”s? a.k.a. N-hot
- Still so sparse?
  - Dense representation?

How to Give a Better Value?

- Categorical meaningful?
  - Frequency/Weights/Information bearing/Topic...
- Overall menaingful?
  - Distributed/Neural/...

Categorical Meaningful Representations

- Term Frequencies

Term	Document						
	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	2	1	0	0	0	0	0
t2	1	2	0	0	0	0	1
t3	3	1	0	0	1	1	0
t4	0	0	1	2	1	1	1
t5	0	0	1	1	1	1	1
t6	0	0	1	1	0	0	0

Topic Representation

	Word 1	Word 2	Word ...	Word N
Topic 1	$P(w_1\|t_1)$	$P(w_2\|t_1)$	...	$P(w_N\|t_1)$
Topic 2	$P(w_1\|t_2)$	$P(w_2\|t_2)$	...	$P(w_N\|t_1)$
Topic 3	$P(w_1\|t_3)$	$P(w_2\|t_3)$	...	$P(w_N\|t_1)$
...	...	...	...	...
Topic K	$P(w_1\|t_K)$	$P(w_2\|t_M)$	...	$P(w_N\|t_1)$

Topic-Word Distribution

Summary

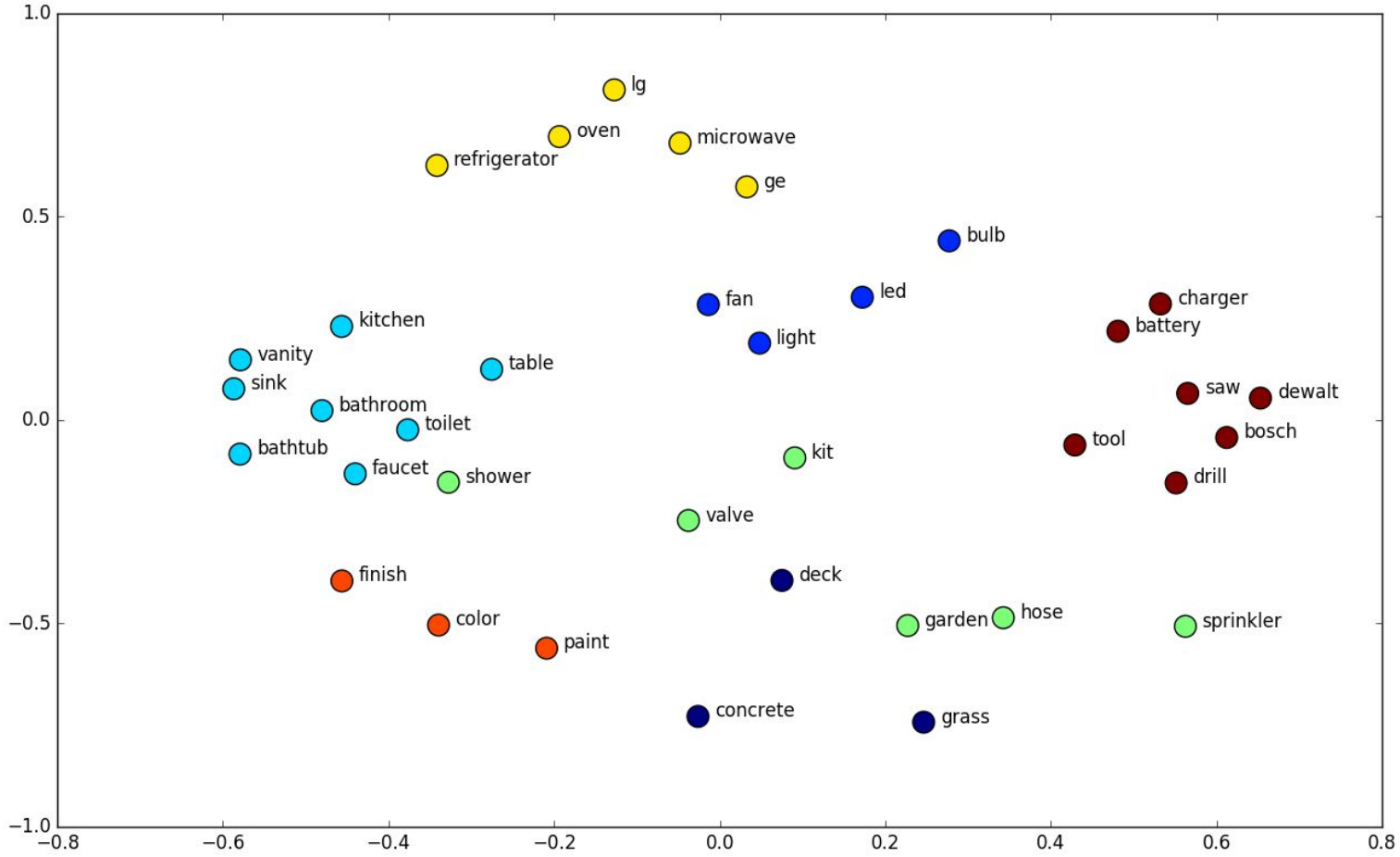
- What can you conclude from the above representations?
  - Intuitive
  - Easy to measure and evaluate
  - Static
  - …
- Are they good enough?

Neural Text Representation

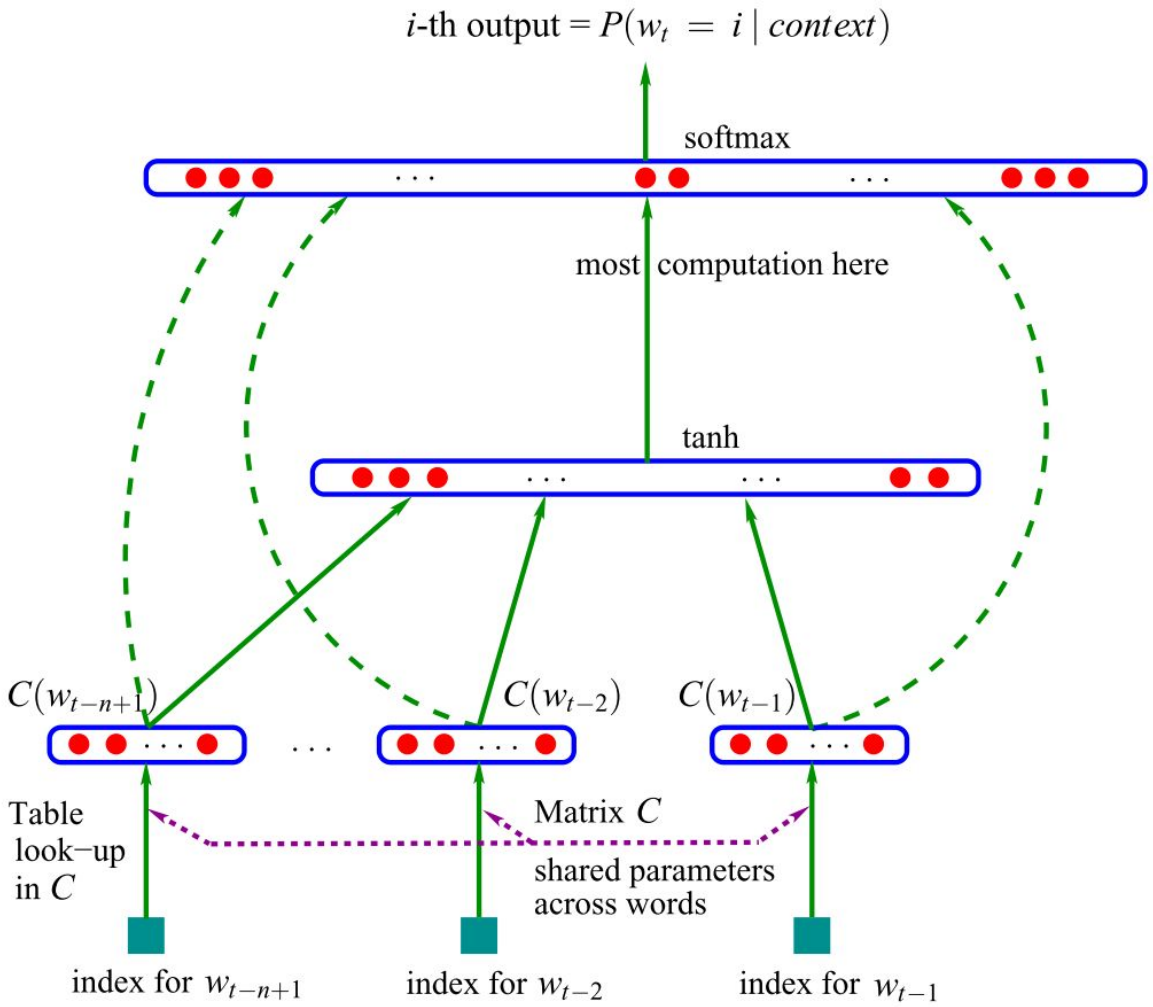
- Word embedding
  - Not intuitive
  - Hard to evaluate
  - Dynamic
  - …
- Why we need that?

$$cat = \begin{bmatrix} 0.4546 \\ -0.1112 \\ 0.8891 \\ -0.3439 \\ -0.7611 \\ 0.5111 \\ 0.4321 \\ 0.9999 \end{bmatrix}$$

Neural Text Representation



Neural Text Representation

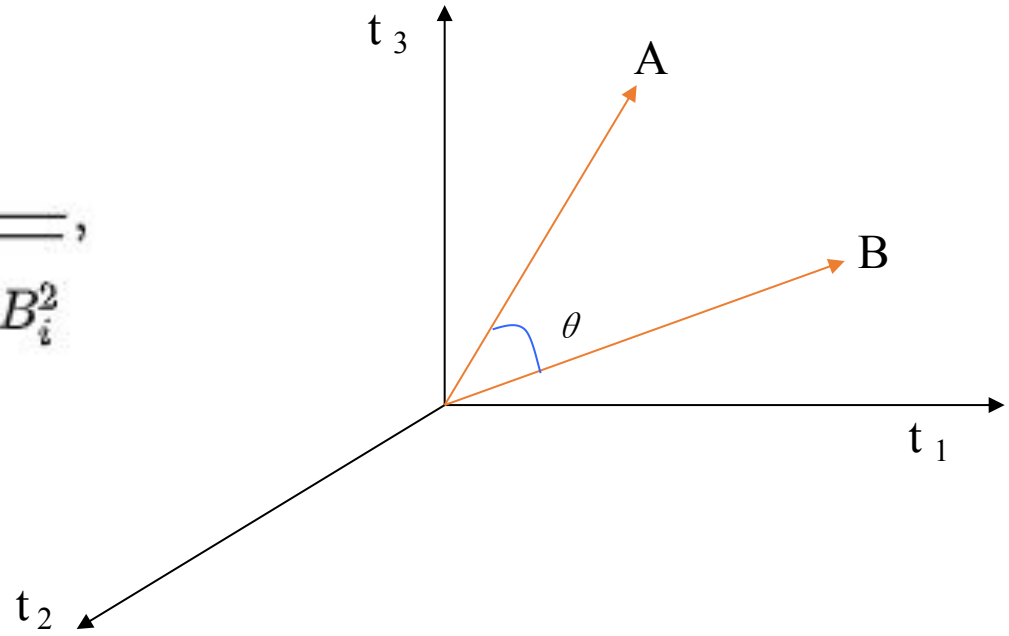


Evaluate?

- How do you know a representation is better?
- Idea: How about measure their relations instead of direct assessment
  - E.g., Euclidean distance? Distance between two v_1 and v_2 : $|v_1 - v_2|$.
- Why this (or similar way) is not a good idea?
  - Serious normalization problem
- How about we look at vector angles?

Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$



- -1, exact opposite
- 0, orthogonality or decorrelation
- 1, exact the same

Summary

- We need text representions is a compromise with computer!
- One-dimensional array is the main format
- There are various ways to represent text
  - Depending on how you want to use it
- Conventional Methods
- What can be done with neural text representations
- How to evaluate

Homework Assignment 1

- Presentations from 3 groups (already in Canvas) for different text representation methods (40 mins)
- Topic per group (difficulties are marked with ★)
  - TF/IDF - Term Frequency / Inversed Document Frequency ★
  - LSA - Latent Semantic Analysis (with SVD) ★★
  - LDA - Latent Dirichlet Allocation (with Gibbs sampling) ★★★
- Policy
  - Topics are “first come first occupy”
  - Topic for each team are different
  - Higher credits will be obtained if you present a difficult one

Homework Assignment 1

- Outlines for your presentation (must have items)
  - Introduction to the method
  - Alogrithm details (architecture, flow, etc...)
  - Performance (showcase)
  - Potential Application
- Requirements
  - At least two students should present
  - Everyone's work in a group should be listed in the presentation
  - NO OVERTIME allowed