

# NATURAL LANGUAGE PROCESSING: ОБЗОР

ДМИТРИЙ НОВИЦКИЙ

1

# ВВЕДЕНИЕ: ЗАДАЧИ, РЕШАЕМЫЕ NLP

- Машинный перевод
- Понимание текста
- Текстовый поиск
- Кластеризация и классификация текстов
- Аннотирование и реферирование
- Генерация текста
- Диалоговые системы (чатботы)

# ТРАДИЦИОННЫЙ (GOOD OLD FASHIONED)


- **Первичный:** токенизация
- **Морфологический:** лемматизация
- **Синтаксический:** Синтаксический анализ, парсинг, синтаксические деревья
- **Семантический:** тезаурусы, семантические сети, онтологии, лексические функции
- **Прагматический:**

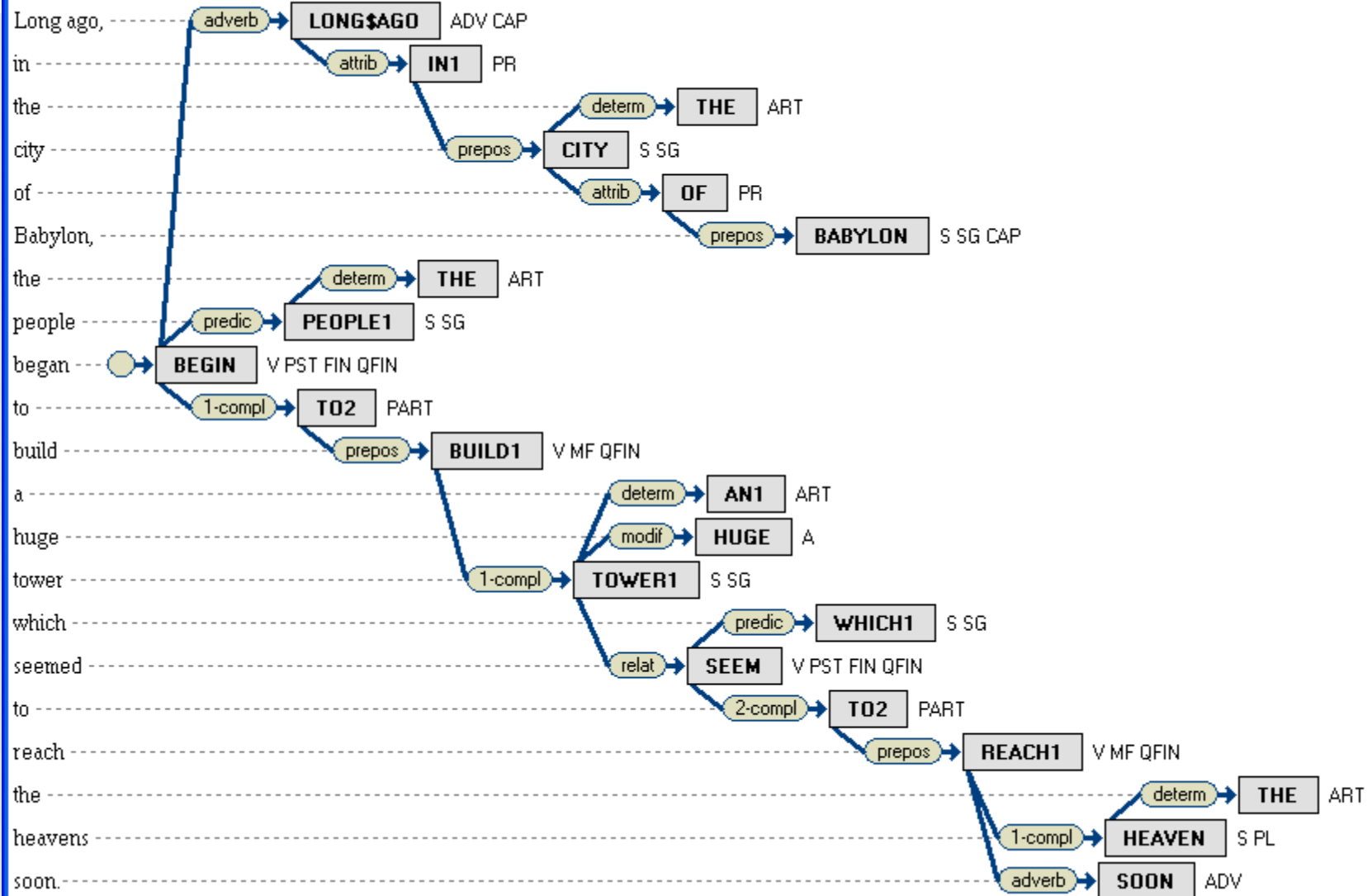
# ЛЕММАТИЗАЦИЯ

## Основные процедуры обработки ЕЯ

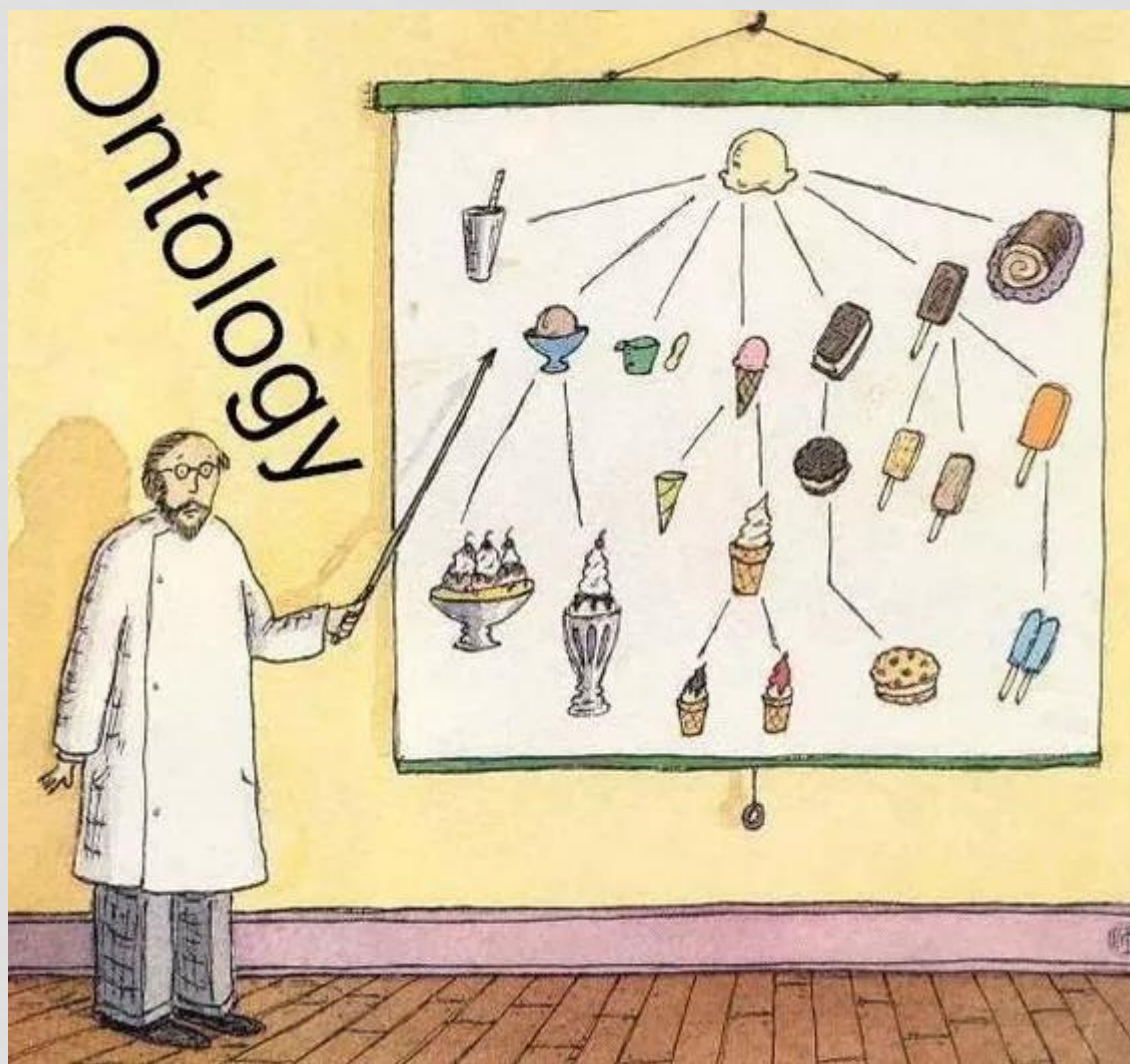
- **Лемматизация** – процесс образования первоначальной формы слова, исходя из других его словоформ.
- Во многих языках слово может встречаться в нескольких формах с различными флексиями.
- Например, английский глагол **walk** может быть представлен следующими формами: **walk**, **walked**, **walks**, **walking**.
- Базовая форма, **walk**, зафиксированная в словаре, называется **леммой слова**.

# СИНТАКСИЧЕСКИЙ АНАЛИЗ

■ Sentence: Long ago, in the city of Babylon, the people began to build a huge tower which seemed to reach the heavens... 

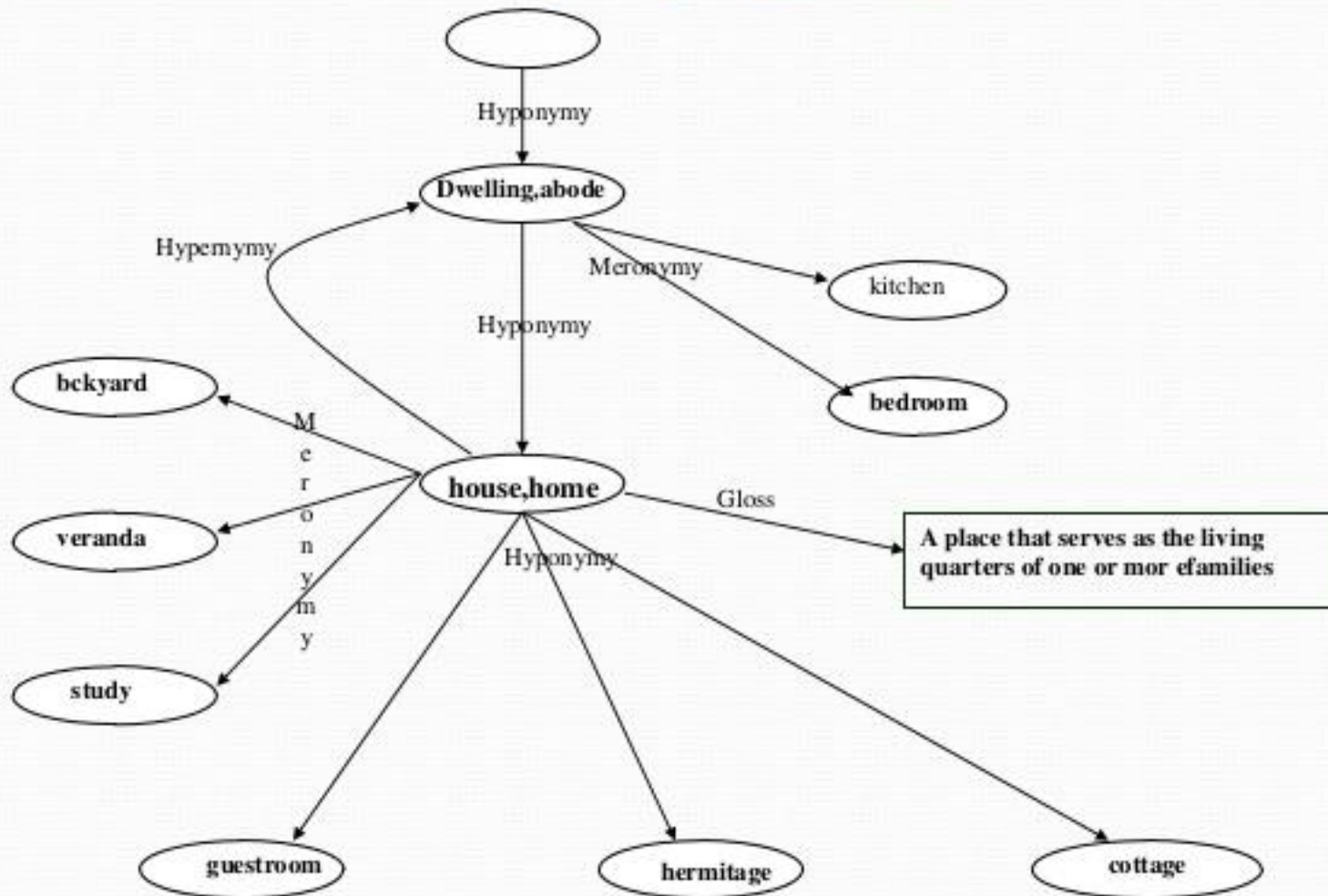


# ОНТОЛОГИИ



# СЕМАНТИЧЕСКАЯ СЕТЬ (WORDNET)

## WordNet Sub-Graph (English)





# ЛЕКСИЧЕСКИЕ ФУНКЦИИ

**Syn** – синоним; **Syn<sub>c</sub>**, **Syn<sub>з</sub>** и **Syn<sub>п</sub>** обозначают, соответственно, синоним с более узким, более широким и пересекающимся значением. (Символы **c**, **з** и **п** используются в том же смысле при **Conv**, **Anti** и некоторых других ЛФ.) Примеры: **Syn**(стрелять) = палить; **Syn<sub>c</sub>**(стрелять) = обстреливать.

**Conv<sub>ij</sub>** – конверсив, т.е. лексическая единица с тем же смыслом, что и **C<sub>0</sub>**, но с перестановкой аргументов *i* и *j*: **Conv<sub>21</sub>**(включать) = принадлежать [множеству]; **Conv<sub>231</sub>**(мнение) = репутация.

**Anti** – антоним: **Anti**(победа) = поражение.

**Gener** – такое родовое понятие, что '**Gener** + **C<sub>0</sub>**' = '**C<sub>0</sub>**' (где **C<sub>0</sub>** – заглавная лексема): **Gener**(газ) = вещество, ср. газообразное вещество = газ.

**Figur** – стандартная метафора для **C<sub>0</sub>**: **Figur**(блокада) = кольцо [кольцо блокады]; **Figur**(туман) = пелена [пелена тумана].

**Dimun** – диминутив, или уменьшительная форма: **Dimun**(дом) = домик; **Dimun**(озеро) = озерко.

**Augm** – аугментатив, или увеличительная форма: **Augm**(дом) = домище, домина; **Augm**(рука) = ручища.

**S<sub>0</sub>**, **A<sub>0</sub>**, **Adv<sub>0</sub>**, **V<sub>0</sub>** – синтаксические дериваты от **C<sub>0</sub>**, т.е., соответственно, существительное, прилагательное, наречие и глагол, имеющие тот же смысл, что и **C<sub>0</sub>**: **S<sub>0</sub>**(стрелять) = стрельба; **A<sub>0</sub>**(стрелять) = стрелковый.

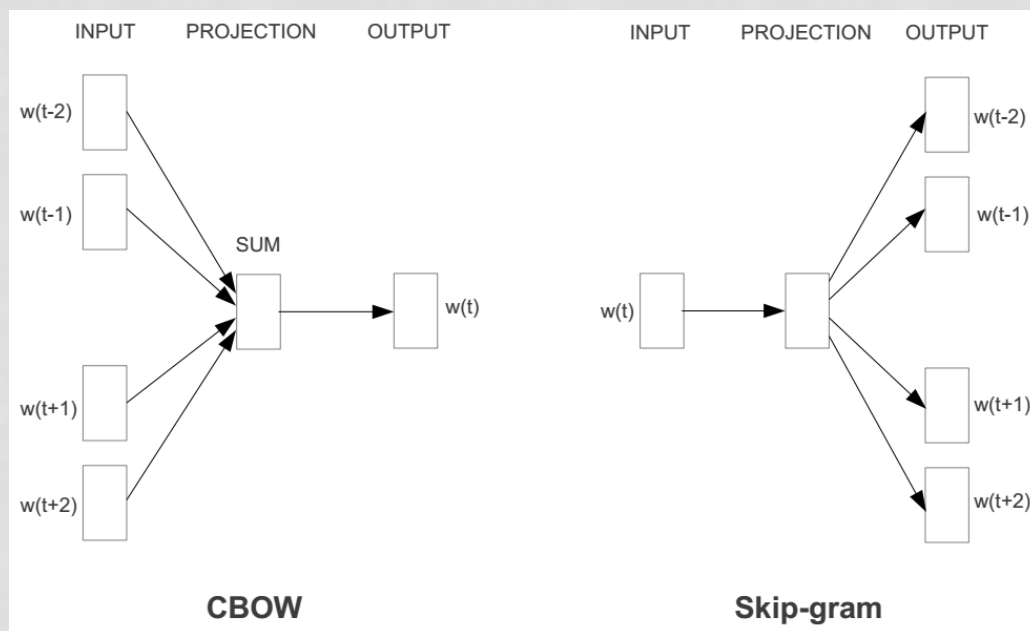


# WORD2VEC : КРАТКОЕ СОДЕРЖАНИЕ

- 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

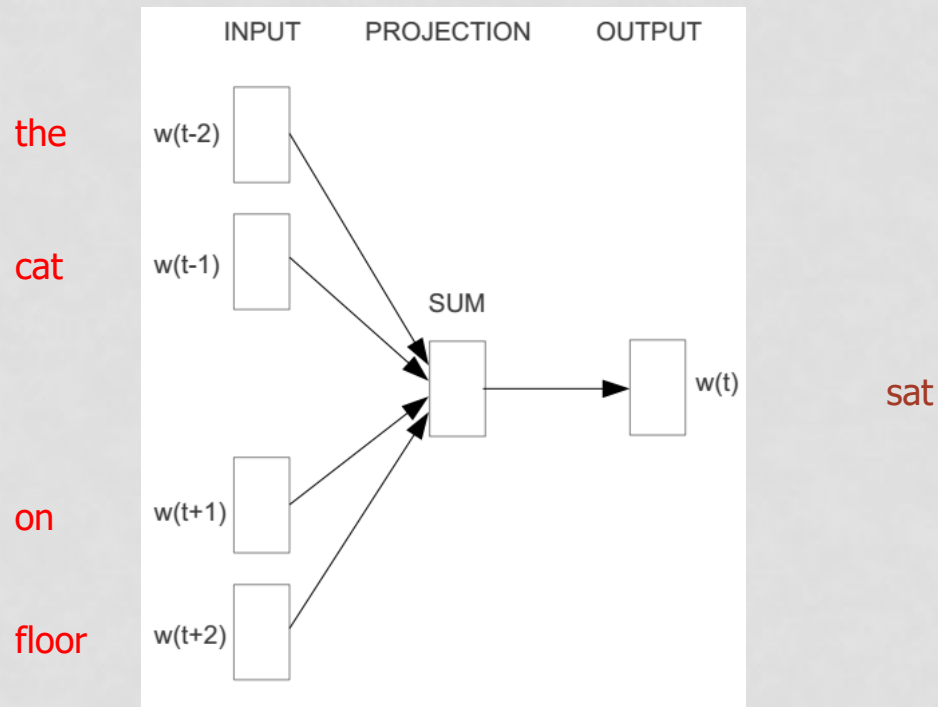
# REPRESENT THE MEANING OF WORD – WORD2VEC

- 2 basic neural network models:
  - Continuous Bag of Word (CBOW): use a window of word to predict the middle word
  - Skip-gram (SG): use a word to predict the surrounding ones in window.



# WORD2VEC – CONTINUOUS BAG OF WORD

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



Index of cat in vocabulary

Input layer



Hidden layer



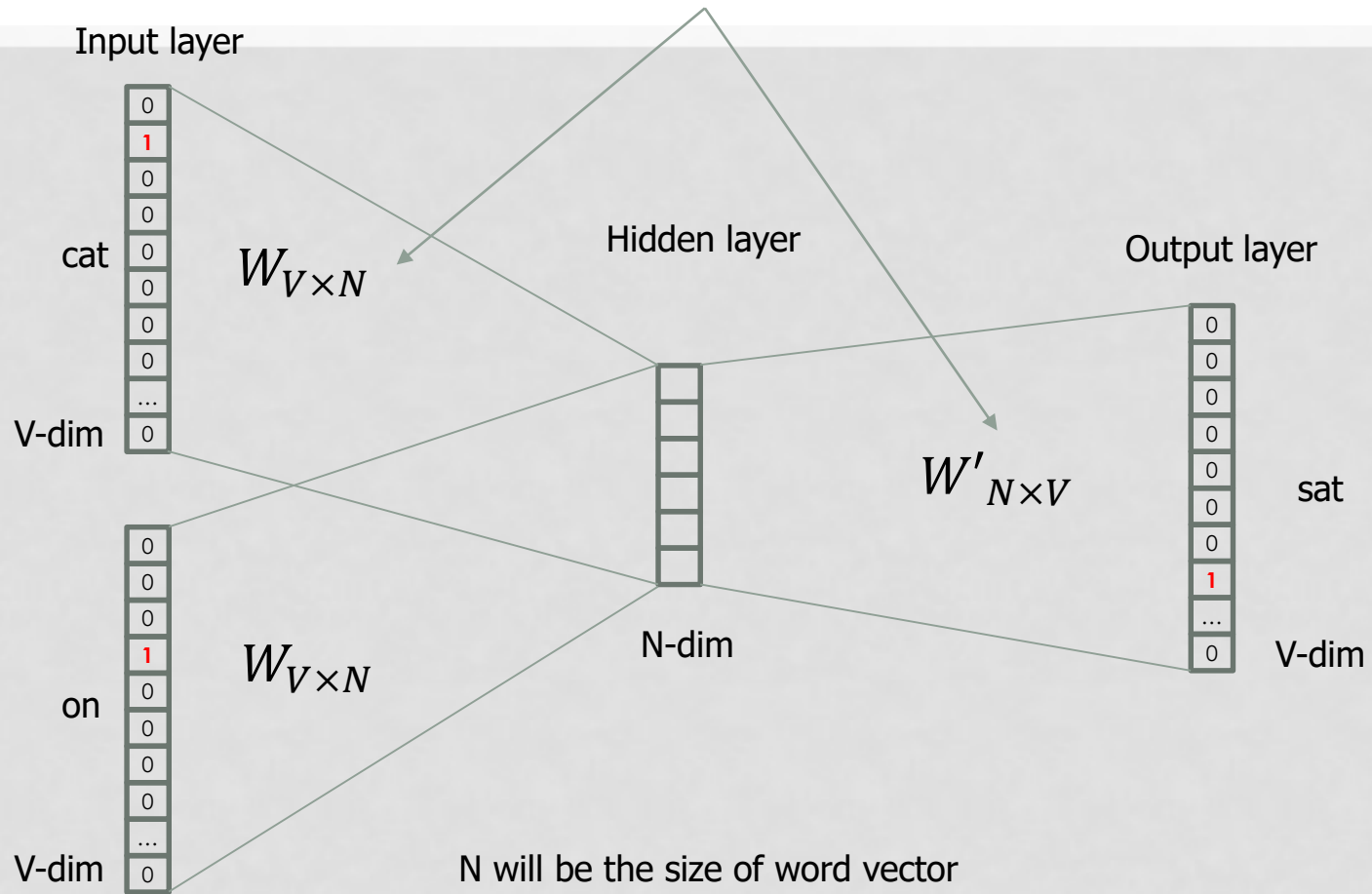
Output layer

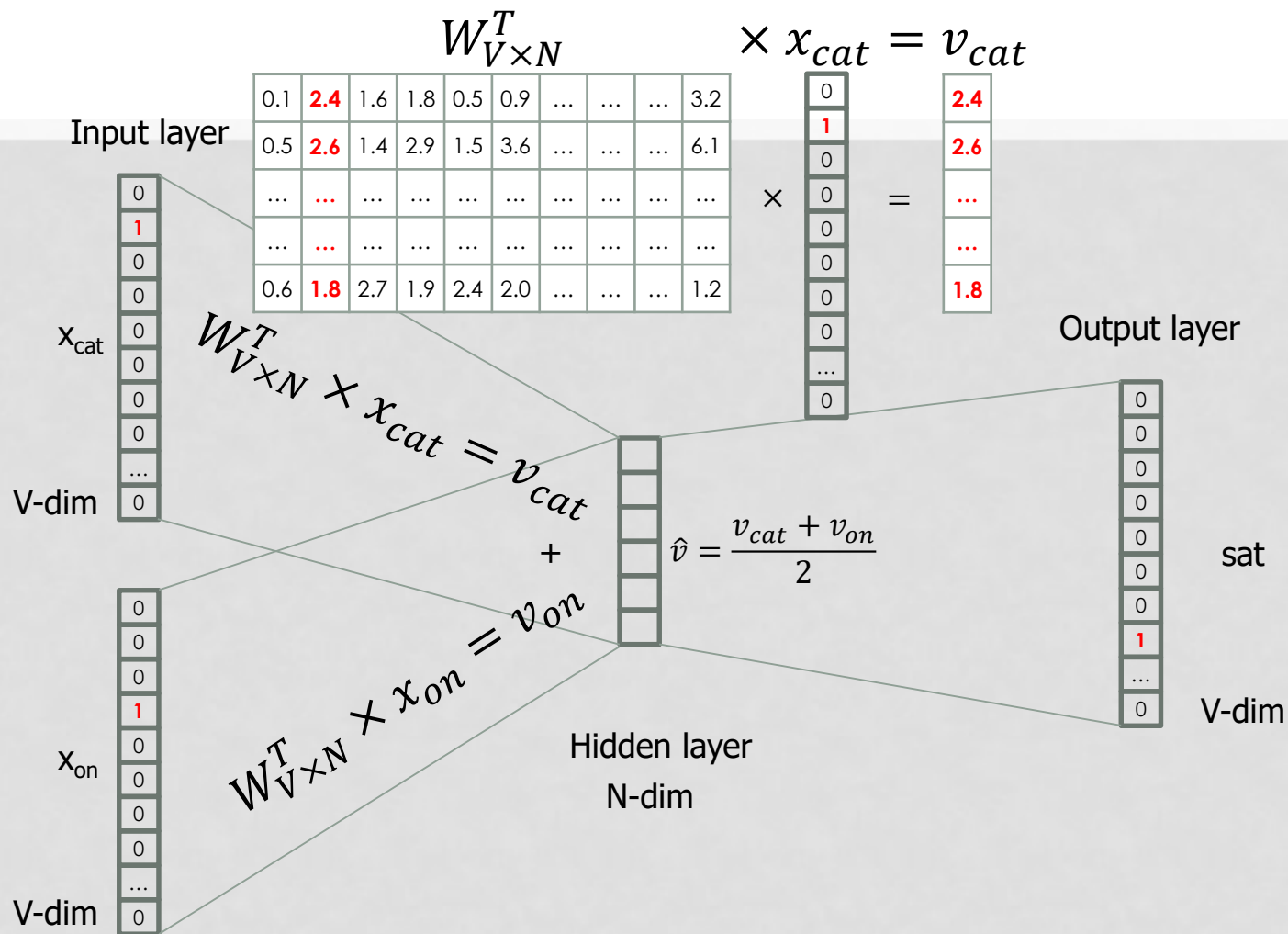


one-hot vector

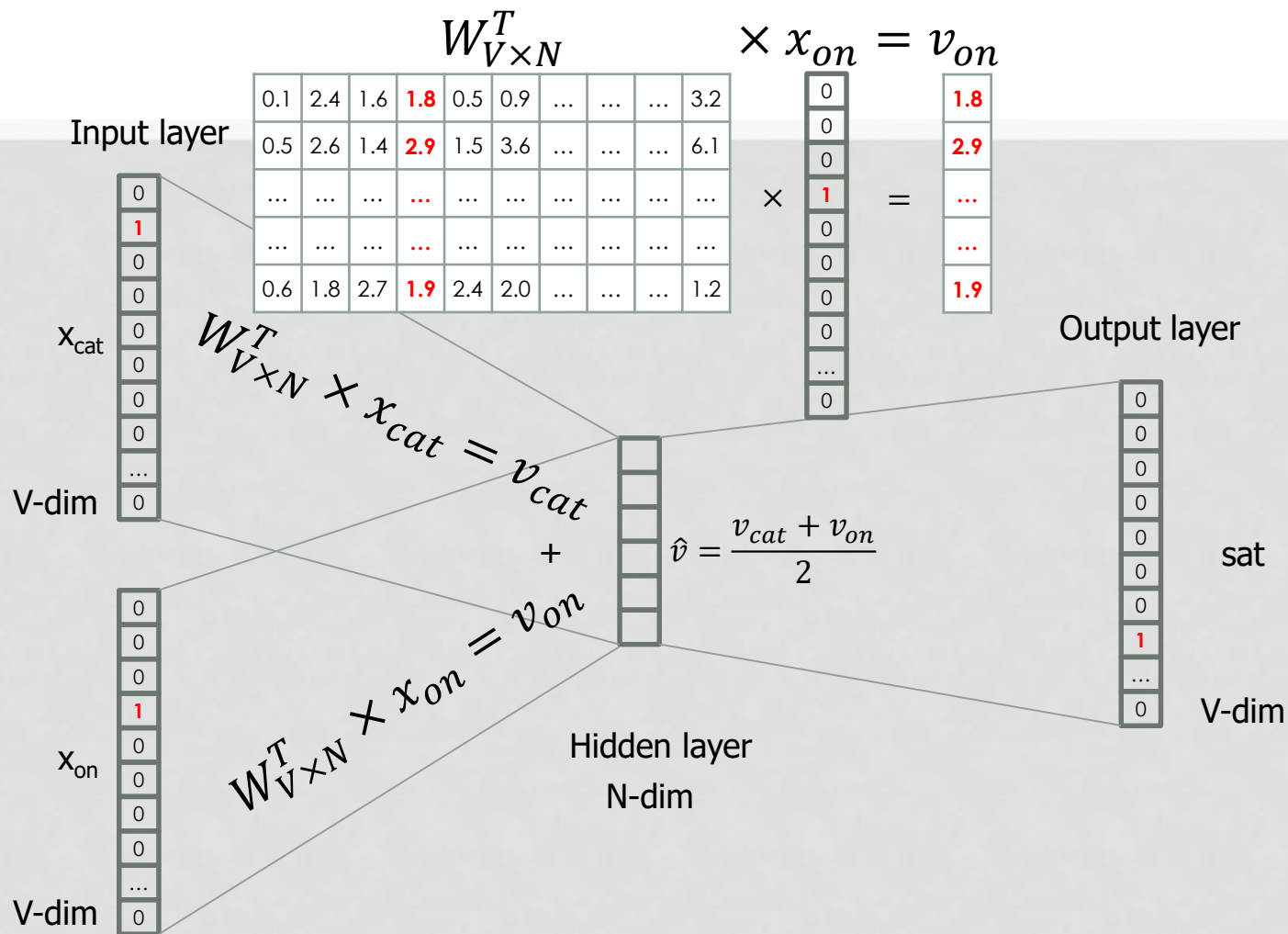
one-hot vector

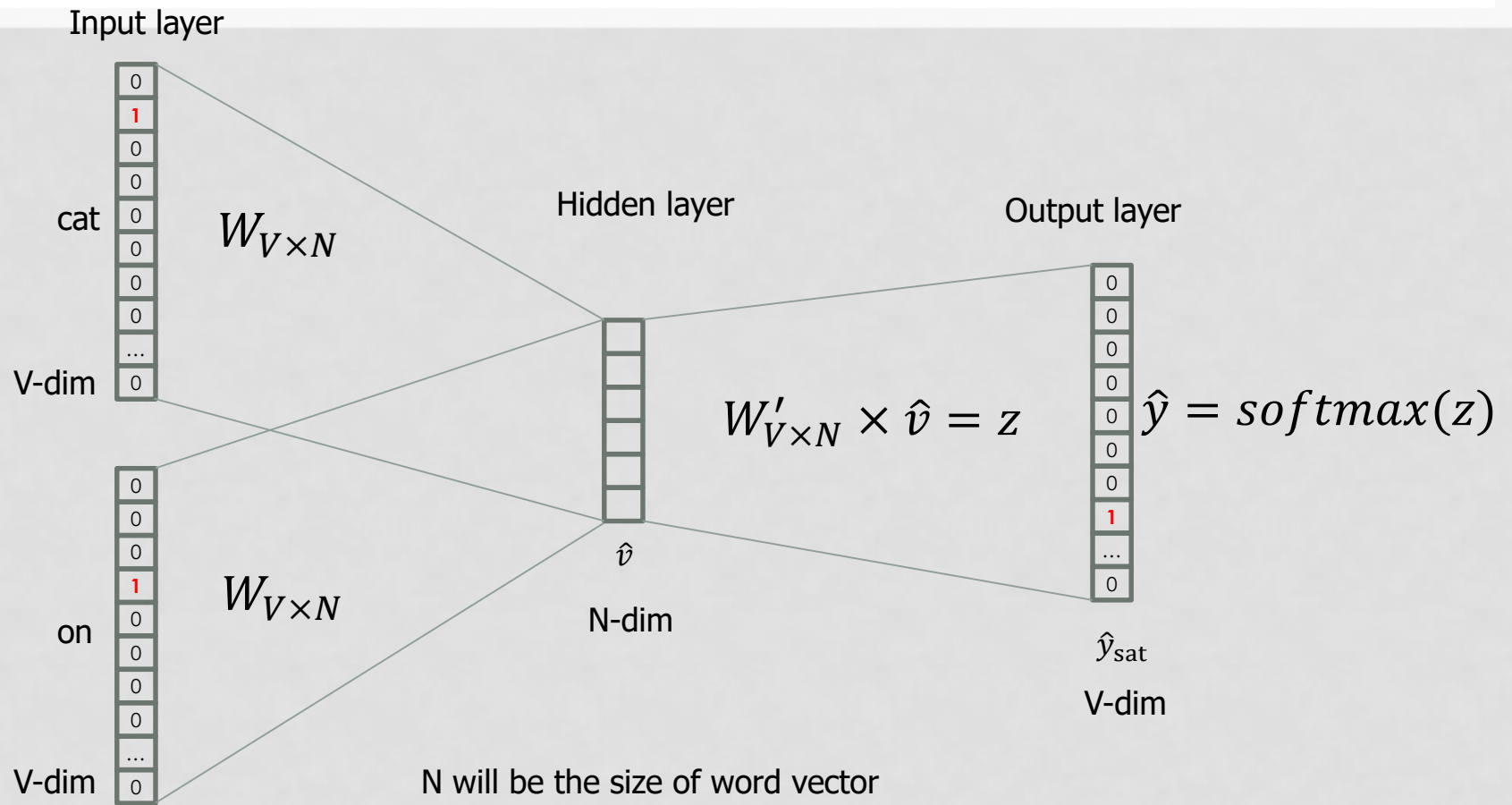
We must learn  $W$  and  $W'$

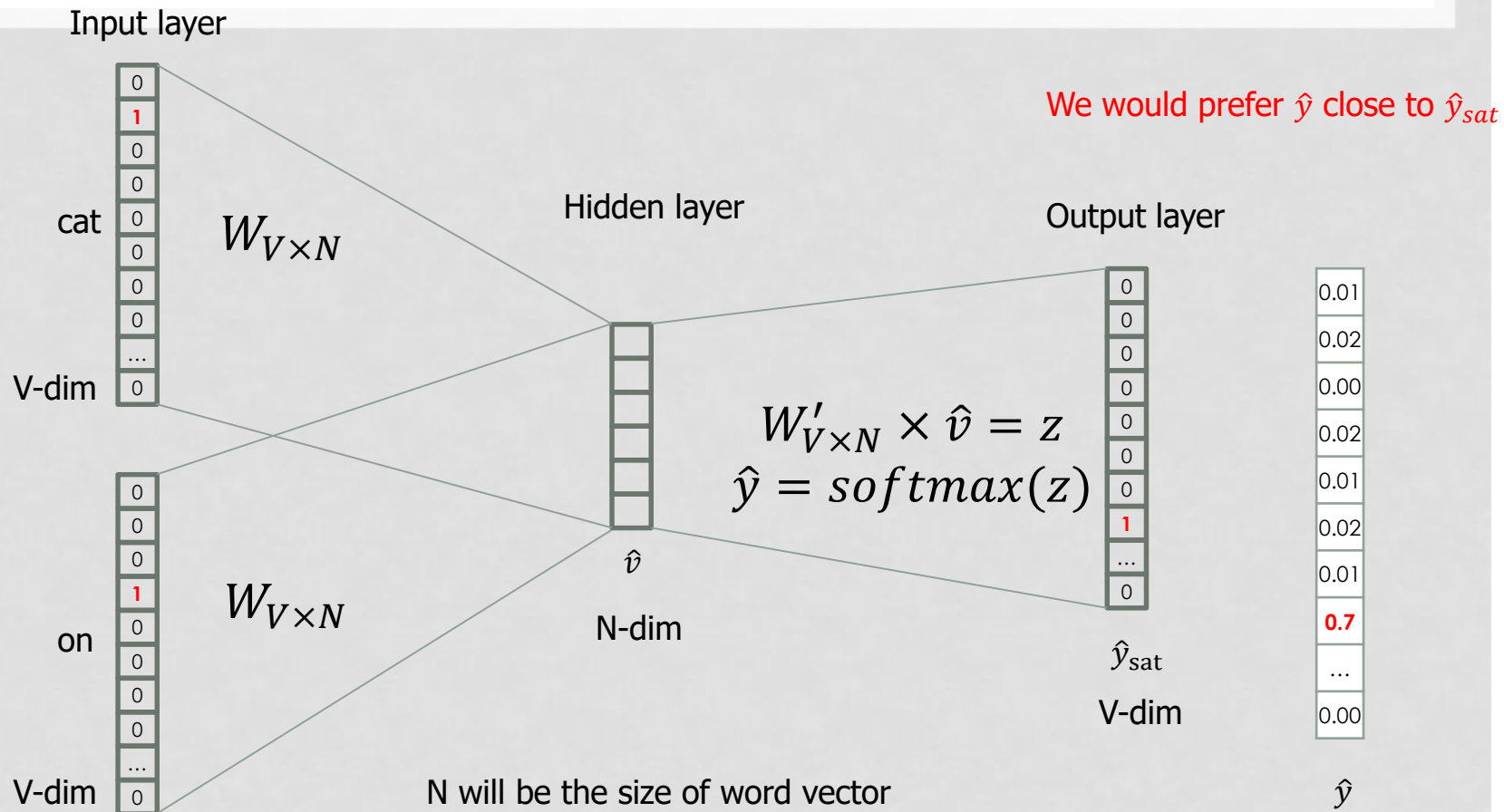


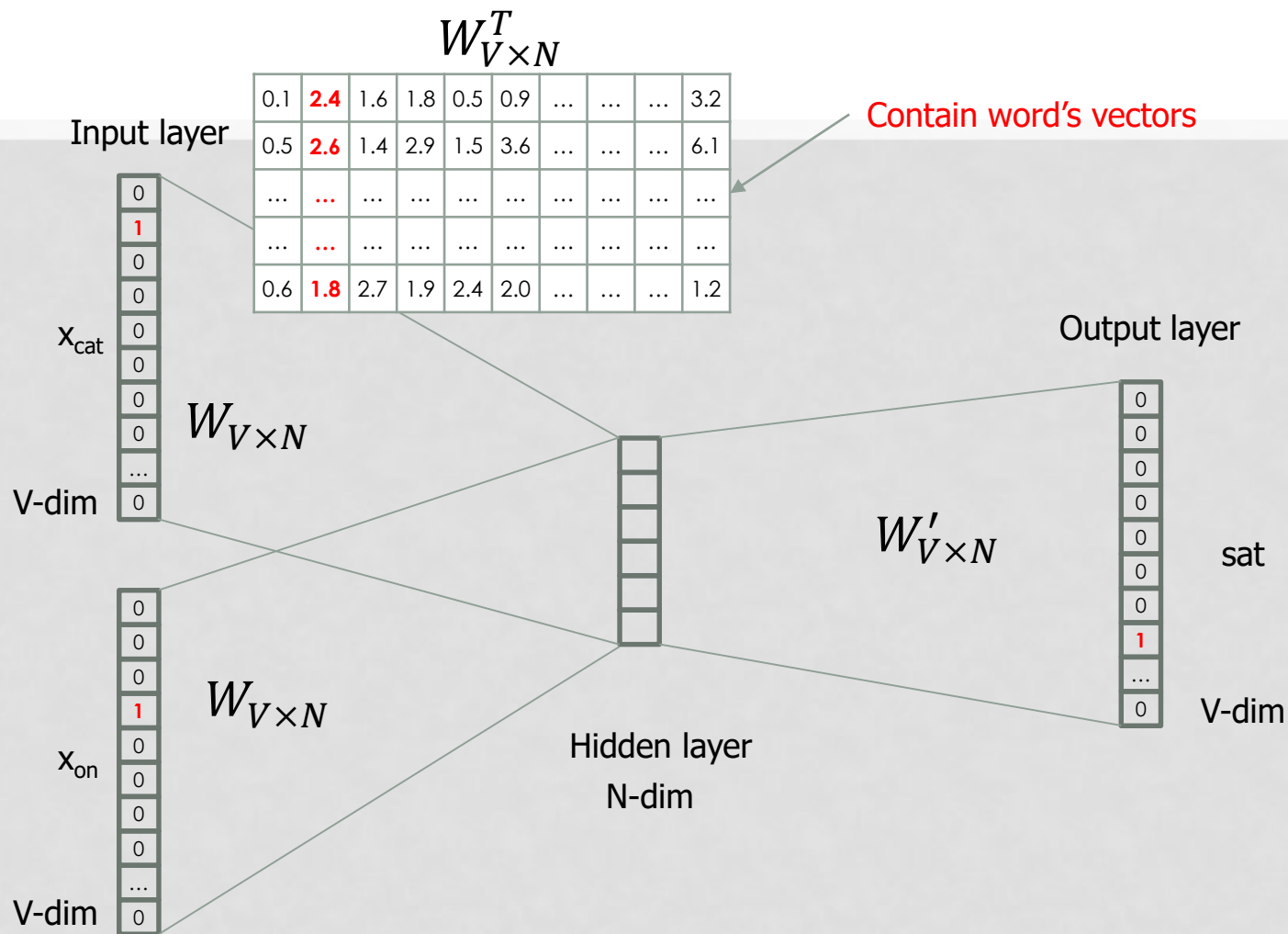












We can consider either  $W$  or  $W'$  as the word's representation.  
Or even take the average.

# SOME INTERESTING RESULTS

## Word Analogies

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

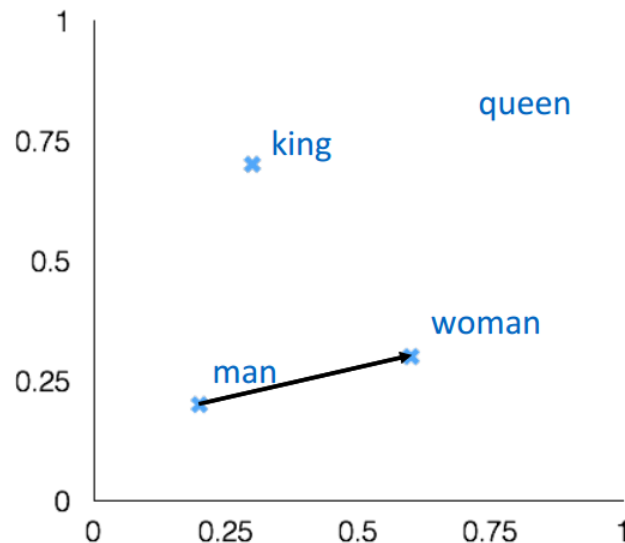
a:b :: c:?



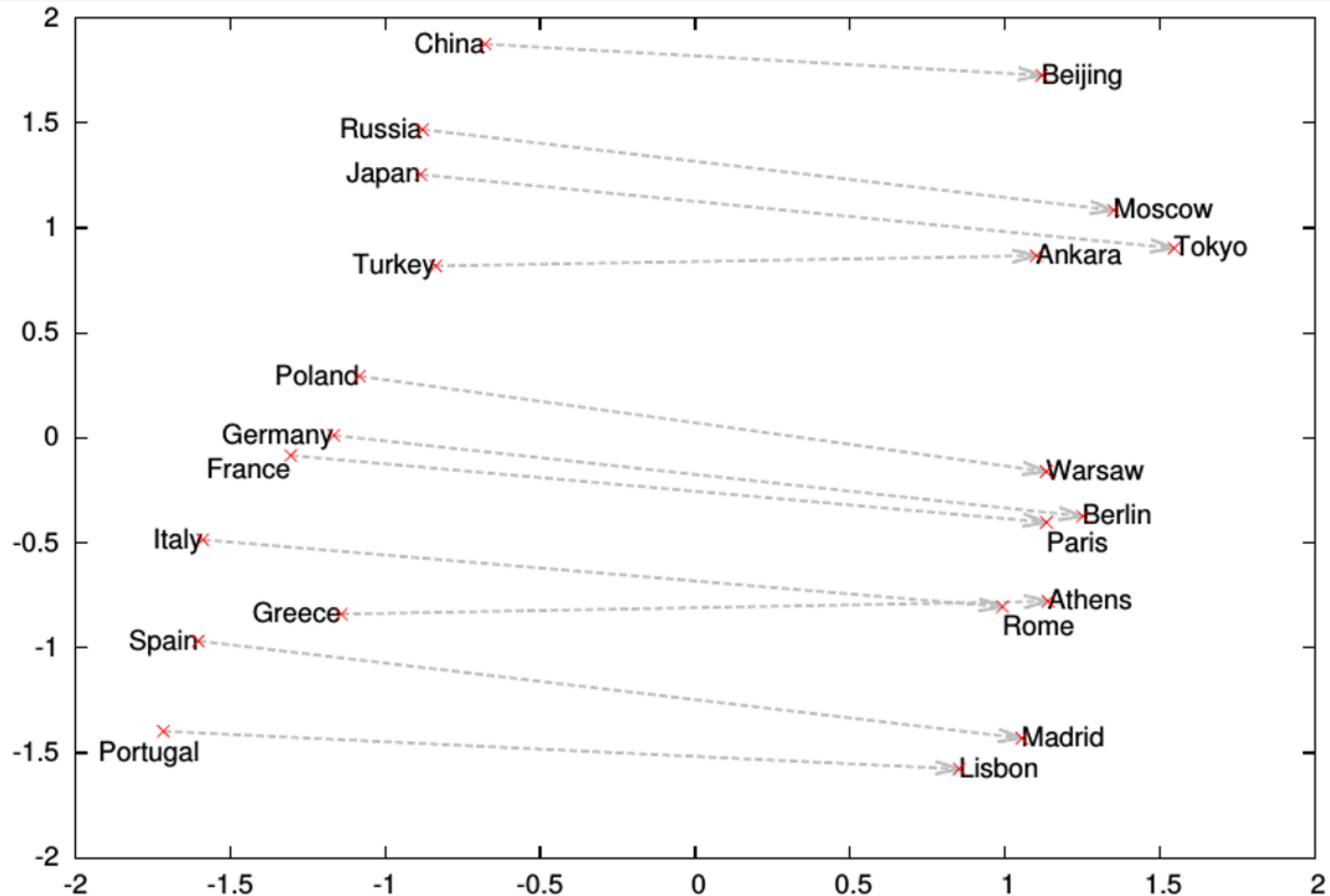
$$d = \arg \max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

+	king	[ 0.30 0.70 ]
-	man	[ 0.20 0.20 ]
+	woman	[ 0.60 0.30 ]
<hr/>		
	queen	[ 0.70 0.80 ]



# WORD ANALOGIES





# ОСНОВНАЯ СТАТЬЯ

- Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics
- Ehsaneddin Asgari,
- Mohammad R. K. Mofrad
- PLOS ONE November 10, 2015
- <https://doi.org/10.1371/journal.pone.0141287>

# РАЗБИВКА БЕЛКОВОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

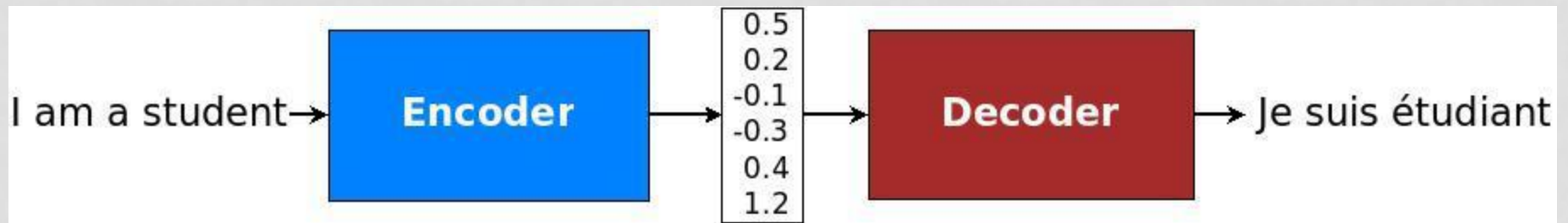
Original Sequence

(1)  $\vec{M}$  (2)  $\vec{A}$  (3)  $\vec{F}$  *SAEDVLKEYDRRRRRMEAL..*

Splittings

{  
1) MAF, SAE, DVL, KEY, DRR, RRM, ..  
2) AFS, AED, VLK, EYD, RRR, RME, ..  
3) FSA ,EDV, LKE, YDR, RRR, MEA, ..

# NEURAL MACHINE TRANSLATION



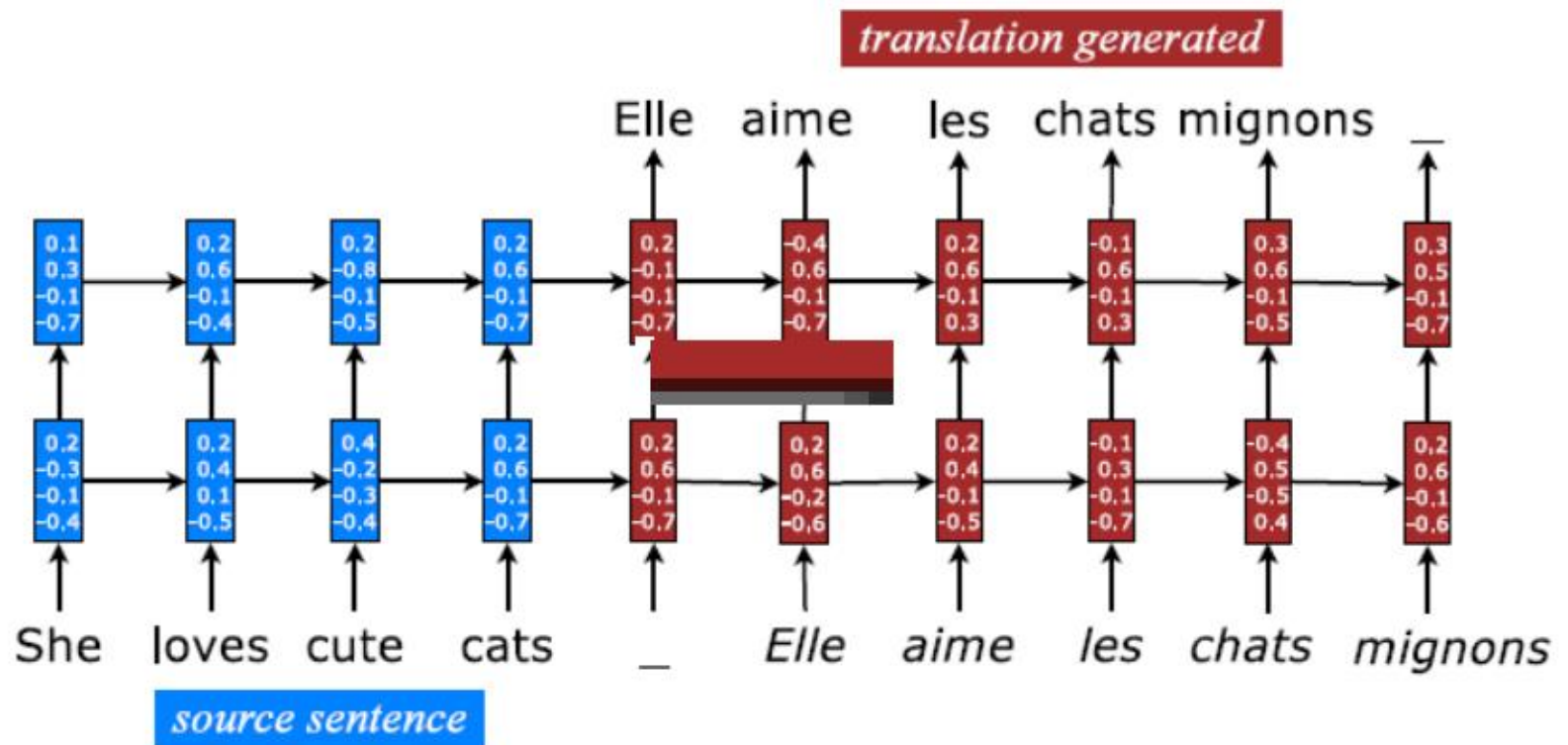
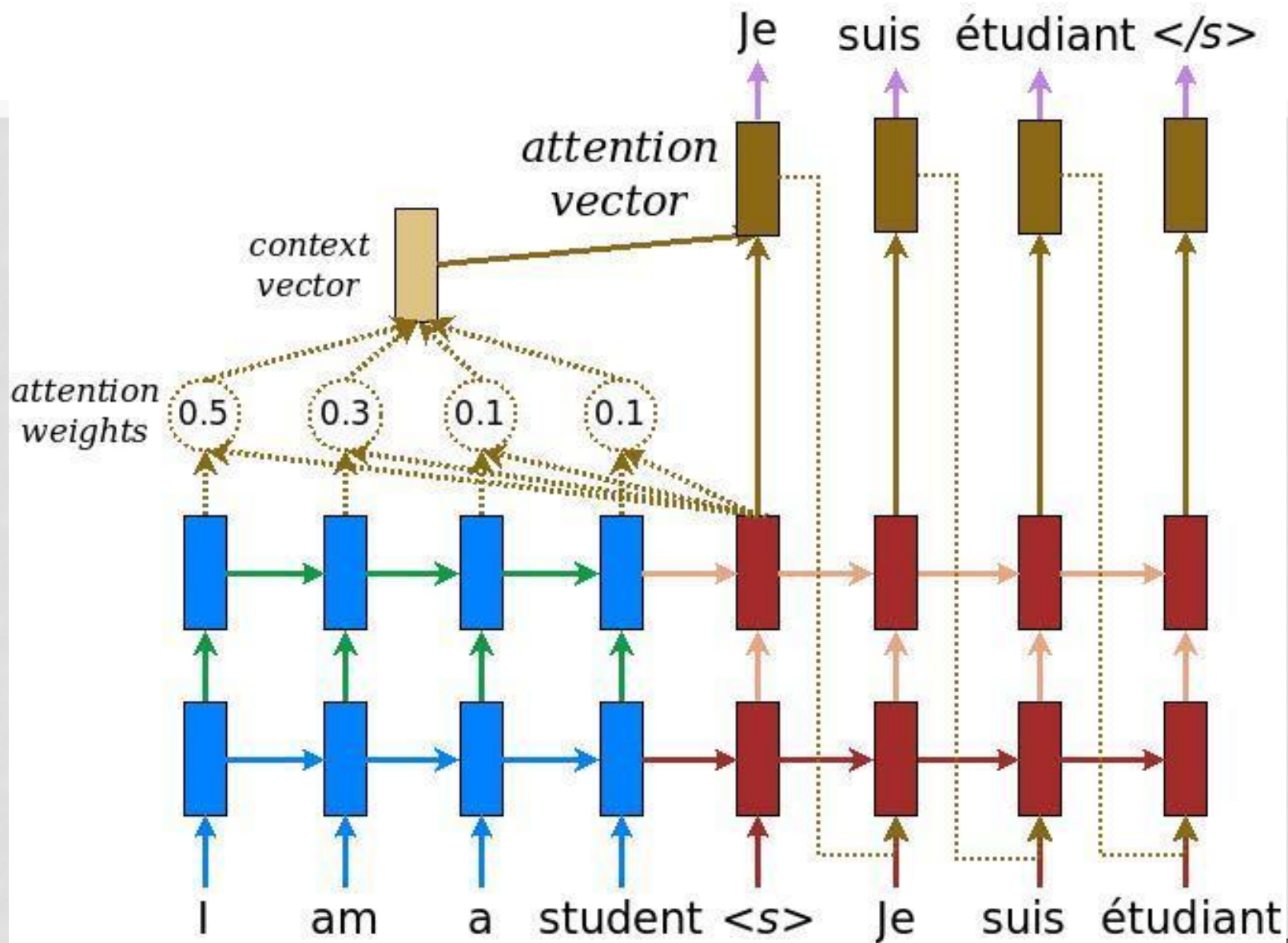


Figure 1.7: **Sequence Models for NMT** – example of a deep recurrent architecture for translating a source sentence “She loves cute cats” into a target sentence “Elle aime les chats mignons”. On the decoder side, *words* generated from previous timesteps are used as inputs for the next ones. Here, “-” marks the end of a sentence.



СПАСИБО ЗА ВНИМАНИЕ!

