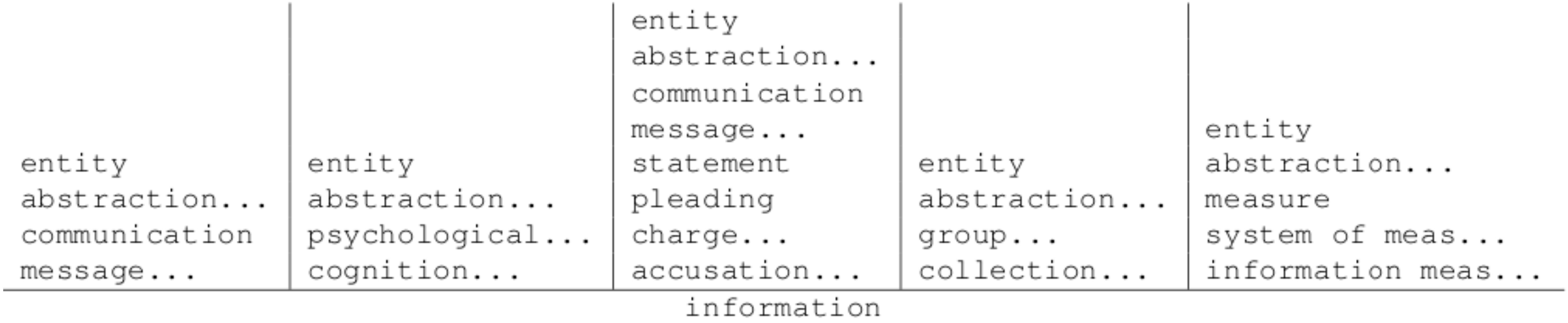
# COMP90042
# Web search and text analysis

Workshop Week 7

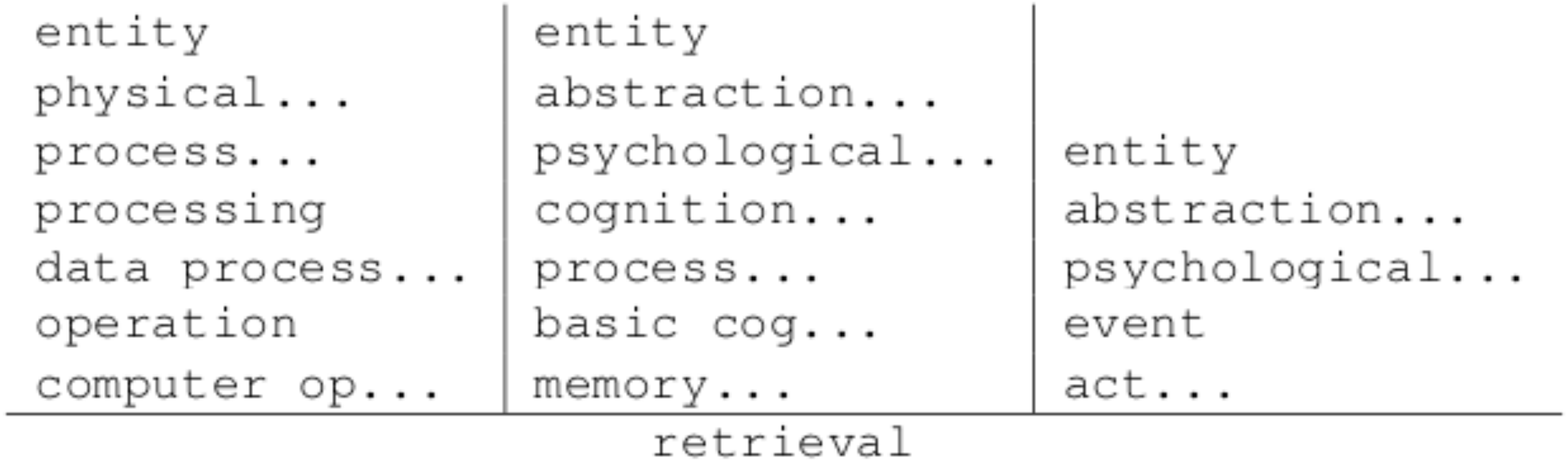
xudong.han@unimelb.edu.au
https://github.com/HanXudong/COMP90042_Workshops

# Review

- Word similarity
- Word embedding



```
                                     entity
                                     abstraction...
                                     communication
                                     message...                            entity
entity           entity              statement         entity              abstraction...
abstraction...   abstraction...      pleading          abstraction...      measure
communication    psychological...    charge...         group...            system of meas...
message...       cognition...        accusation...     collection...       information meas...
------------------------------------------------------------------------------------------
                                        information
```



```
entity               entity
physical...          abstraction...
process...           psychological...    entity
processing           cognition...        abstraction...
data process...      process...          psychological...
operation            basic cog...        event
computer op...       memory...           act...
-------------------------------------------------------
                     retrieval
```

*information* is more similar to the word *retrieval* or the word *science*

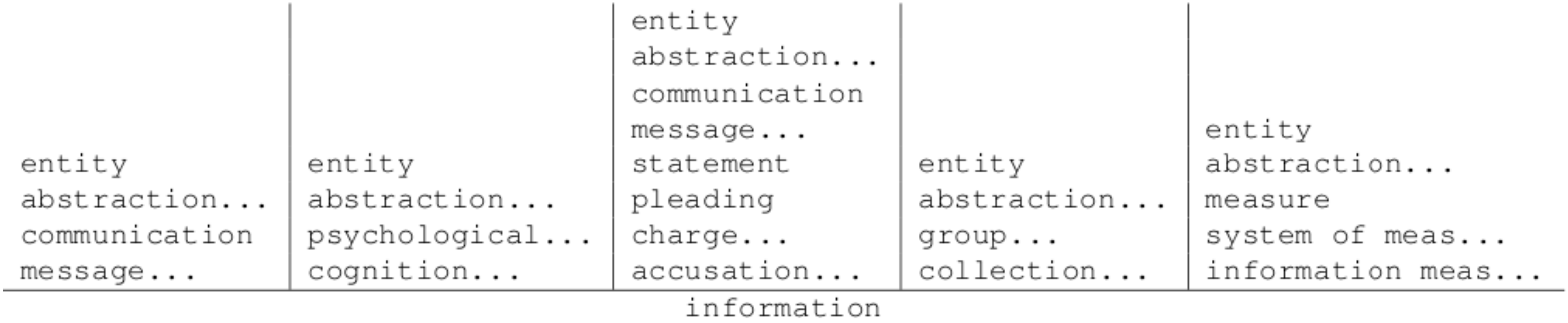
$$WuP\_sim(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

| | | information | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| retrieval | 1 | 0.154 | 0.154 | 0.118 | 0.154 | 0.143 |
| | 2 | 0.308 | 0.615 | 0.235 | 0.308 | 0.286 |
| | 3 | 0.364 | 0.545 | 0.267 | 0.364 | 0.333 |

http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=&o8=1&o1=&o7=&o5=&o9=&o6=&o3=&o4=&r=1&s=science&i=31&h=1000112312312312312312312302220010#c

| | |
|---|---|
| entity | |
| abstraction | |
| psychological | entity |
| cognition | abstraction |
| content | psychological |
| knowledge domain | cognition |
| discipline | ability |

**Science**



| | | | | |
|---|---|---|---|---|
| | | entity | | |
| | | abstraction... | | |
| | | communication | | |
| | | message... | | entity |
| entity | entity | statement | entity | abstraction... |
| abstraction... | abstraction... | pleading | abstraction... | measure |
| communication | psychological... | charge... | group... | system of meas... |
| message... | cognition... | accusation... | collection... | information meas... |

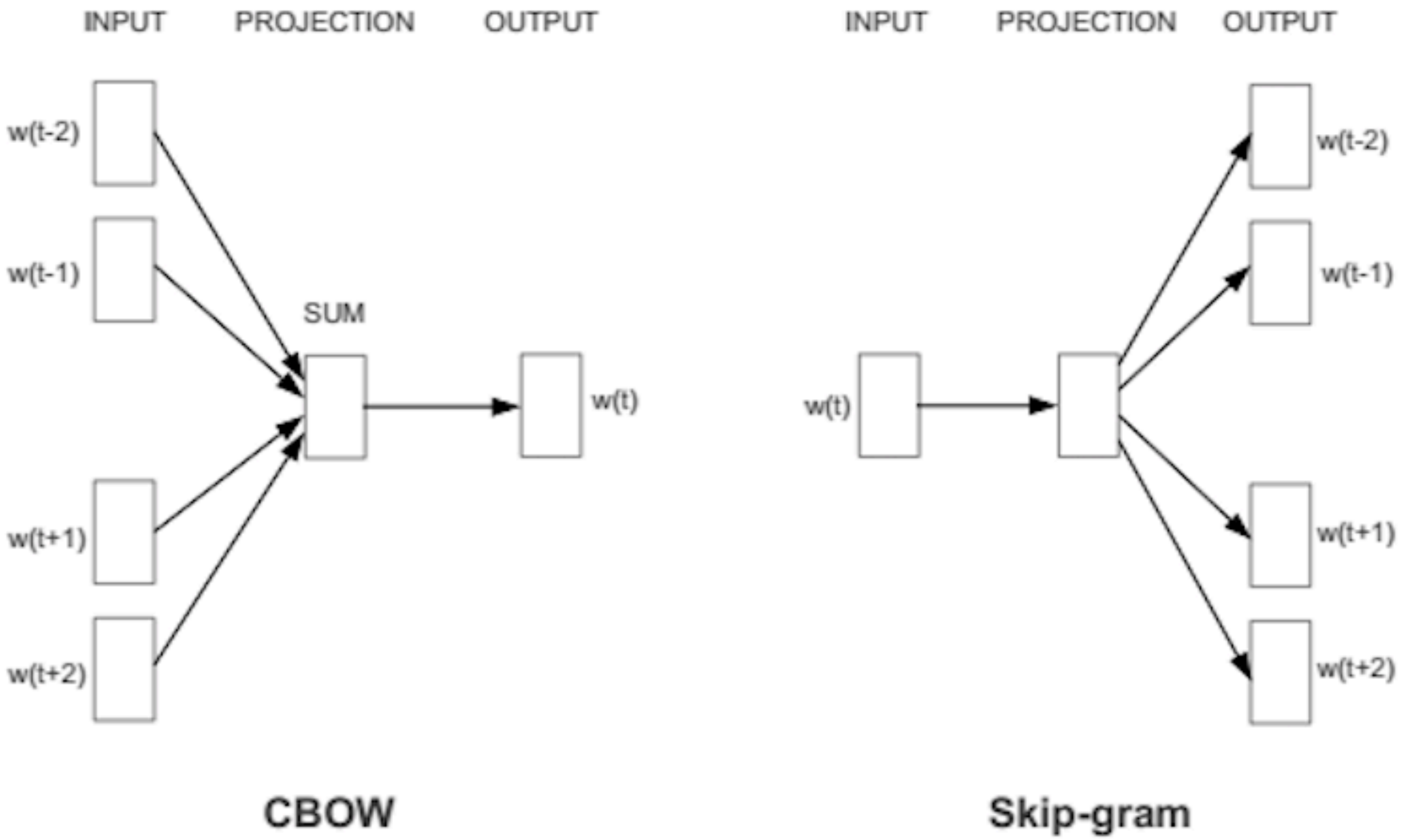
information

| | | information | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| science | 1 | 0.30 | 0.61 | 0.23 | 0.30 | 0.28 |
| | 2 | 0.36 | 0.72 | 0.27 | 0.36 | 0.33 |

# word to vector

https://playground.tensorflow.org/

https://skymind.ai/wiki/word2vec



# At the end of this tutorial you will be able to…

1. explain the main ideas of several common POS tagging approaches

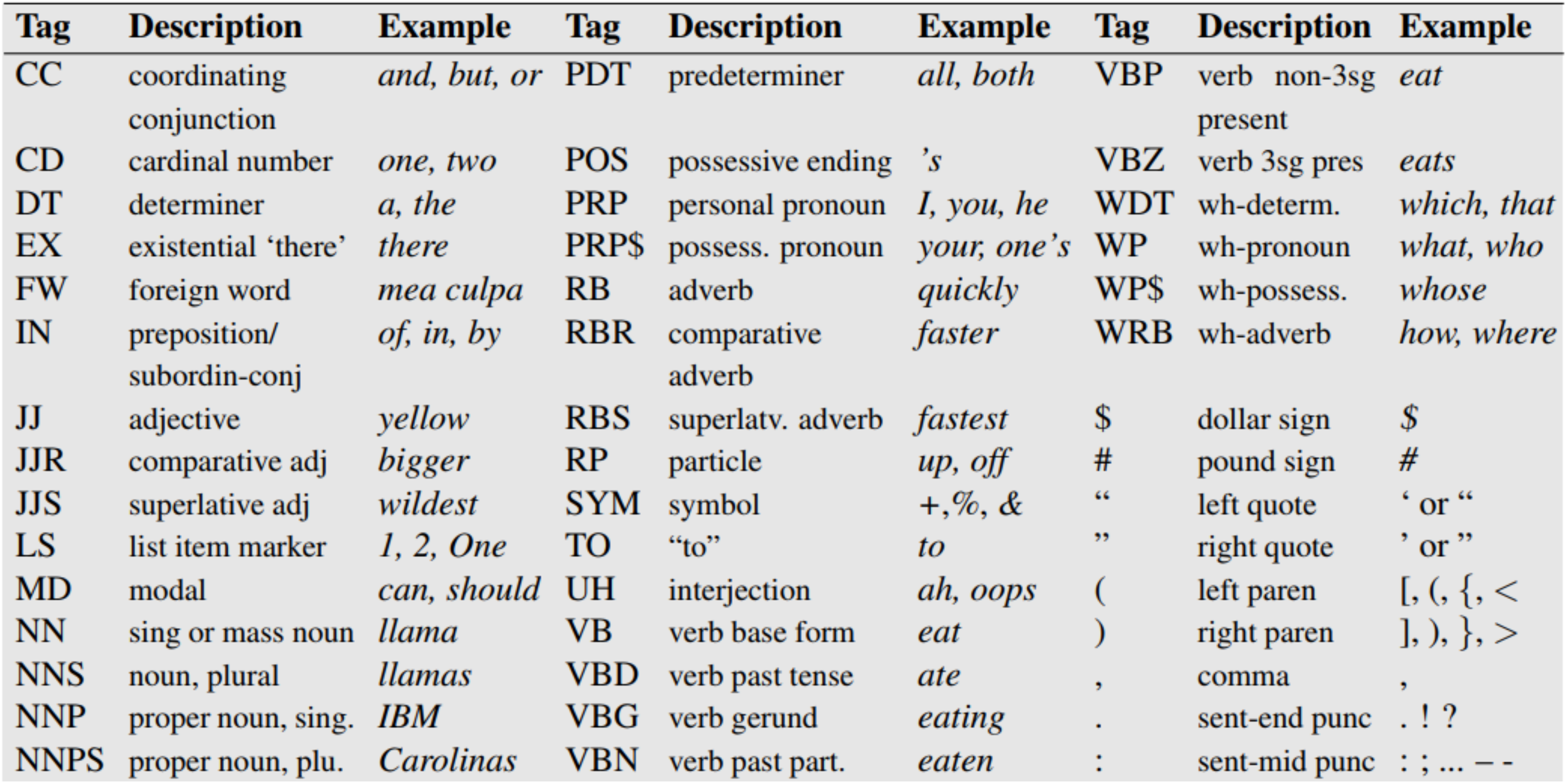
2. do POS tagging manually

3. tell the key differences and similarities between N-gram language model and feed-forward neural language models.

4. explain the basic meaning of RNN and its advantage over the feed-forward model.

# Q1 What is a POS tag

- A part of speech' (abbreviated form: PoS or POS) is a category of words (or, more generally, of lexical items) which have similar grammatical properties.



| Tag | Description | Example | Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|---|---|---|
| CC | coordinating conjunction | *and, but, or* | PDT | predeterminer | *all, both* | VBP | verb non-3sg present | *eat* |
| CD | cardinal number | *one, two* | POS | possessive ending | *’s* | VBZ | verb 3sg pres | *eats* |
| DT | determiner | *a, the* | PRP | personal pronoun | *I, you, he* | WDT | wh-determ. | *which, that* |
| EX | existential ‘there’ | *there* | PRP$ | possess. pronoun | *your, one’s* | WP | wh-pronoun | *what, who* |
| FW | foreign word | *mea culpa* | RB | adverb | *quickly* | WP$ | wh-possess. | *whose* |
| IN | preposition/ subordin-conj | *of, in, by* | RBR | comparative adverb | *faster* | WRB | wh-adverb | *how, where* |
| JJ | adjective | *yellow* | RBS | superlatv. adverb | *fastest* | $ | dollar sign | *$* |
| JJR | comparative adj | *bigger* | RP | particle | *up, off* | # | pound sign | *#* |
| JJS | superlative adj | *wildest* | SYM | symbol | +,%, & | “ | left quote | ‘ or “ |
| LS | list item marker | *1, 2, One* | TO | “to” | *to* | ” | right quote | ’ or ” |
| MD | modal | *can, should* | UH | interjection | *ah, oops* | ( | left paren | [, (, {, < |
| NN | sing or mass noun | *llama* | VB | verb base form | *eat* | ) | right paren | ], ), }, > |
| NNS | noun, plural | *llamas* | VBD | verb past tense | *ate* | , | comma | , |
| NNP | proper noun, sing. | *IBM* | VBG | verb gerund | *eating* | . | sent-end punc | . ! ? |
| NNPS | proper noun, plu. | *Carolinas* | VBN | verb past part. | *eaten* | : | sent-mid punc | : ; ... – - |

# Tagged text Example

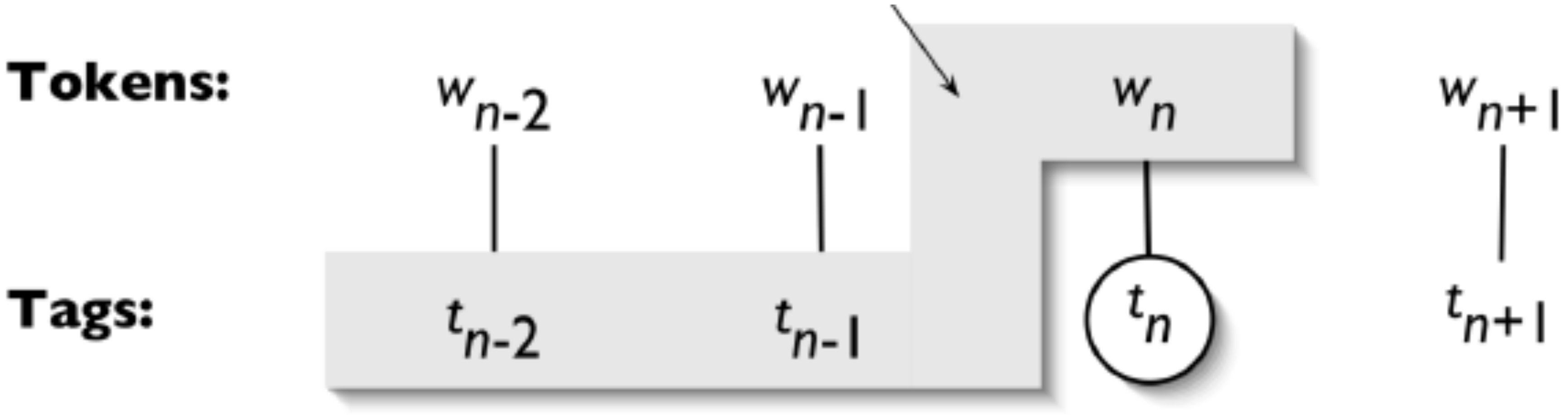
The/DT limits/NNS to/TO legal/JJ absurdity/NN stretched/VBD another/DT notch/NN this/DT week/NN when/WRB the/DT Supreme/NNP Court/NNP refused/VBD to/TO hear/VB an/DT appeal/VB from/IN a/DT case/NN that/WDT says/VBZ corporate/JJ defendants/NNS must/MD pay/VB damages/NNS even/RB after/IN proving/VBG that/IN they/PRP could/MD not/RB possibly/RB have/VB caused/VBN the/DT harm/NN ./.

https://spacy.io/usage/linguistic-features

# Q1a

## What are some common approaches to POS tagging? What aspects of the data might allow us to predict POS tags systematically?

- N-gram
- Rule-based
- Classifier
- HMM



# Q1b

## Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pierre** | **Vinken** | **,** | **61** | **years** | **old** | **,** | **will** | **join** | **the** | **board** | **as** | **a** | **nonexecutive** | **director** | **Nov** | **29** | **.** |
| | | , | | | | , | | | | | | | | | | | . |

| | | |
|---|---|---|
| NN | sing or mass noun | *llama* |
| NNS | noun, plural | *llamas* |
| NNP | proper noun, sing. | IBM |
| VB | verb base form | *eat* |
| JJ | adjective | *yellow* |
| MD | modal | *can, should* |
| CD | cardinal number | *one, two* |
| DT | determiner | *a, the* |
| IN | preposition/ subordin-conj | *of, in, by* |

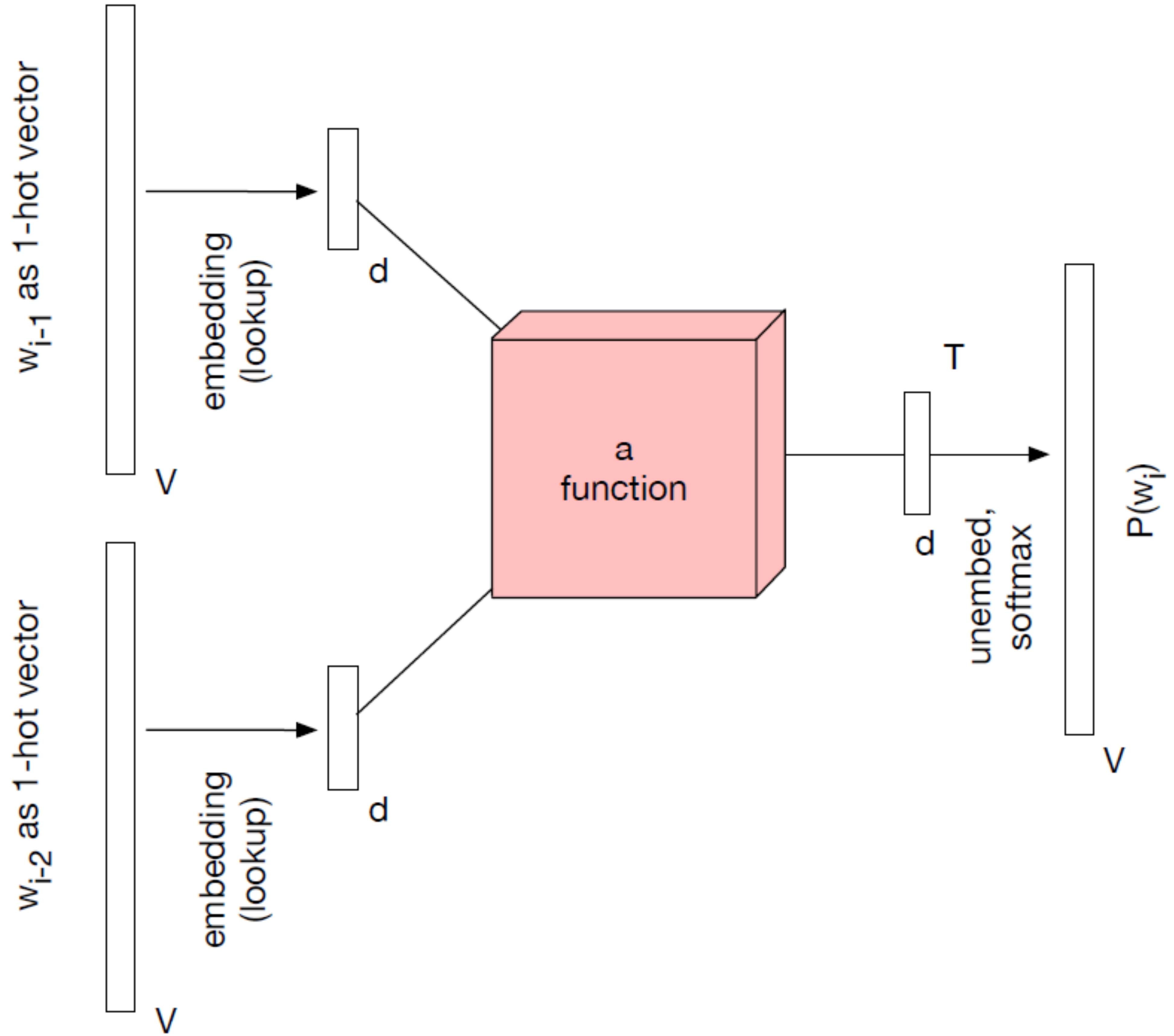
# Q2

## Name the key differences and similarities between n-gram language models versus feed-forward neural language models.

$$P_{add1}(w_i \,|\, w_{i-2}\ w_{i-1}) = \frac{C(w_{i-2}\ w_{i-1}\ w_i) + 1}{C(w_{i-2}\ w_{i-1}) + V}$$

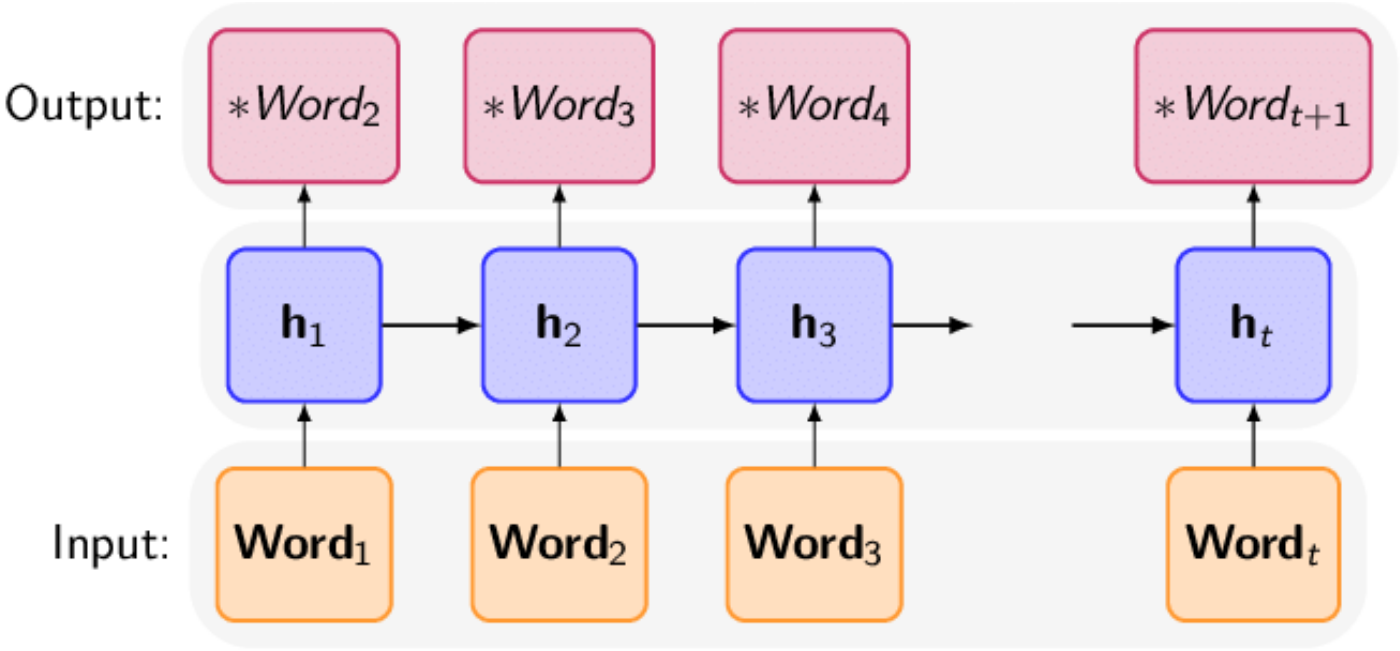
$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \,|\, w_{i-2}\ w_{i-1})$$

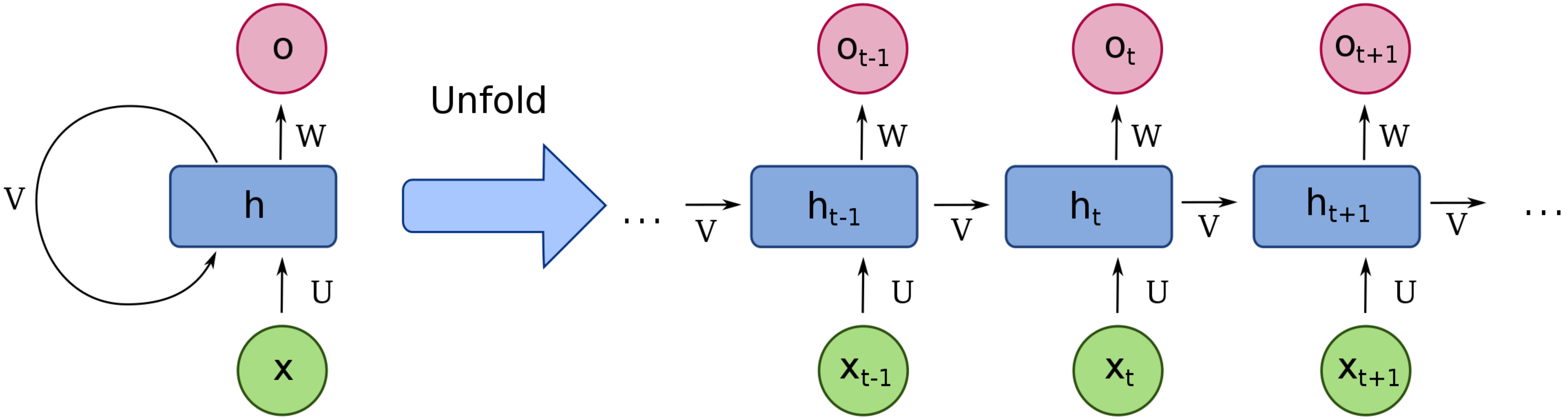
# Feed forward neural net LM



## Q3

## What does recurrent mean in the context of a recurrent neural network (RNN) language model? How does the approach differ from a feed-forward language model?





# Q4
# What advantage does a RNN language model have over a feed-forward language model?

- RNNLM can capture long-distance dependencies, while FFLM cannot. For example, it can balance quotes and brackets over long distances.

- ( ..... ( …. (..) …. ) ….. )