Language models in Supportive Augmentative and Alternative Communication

Shiran Dudy
December, 2019

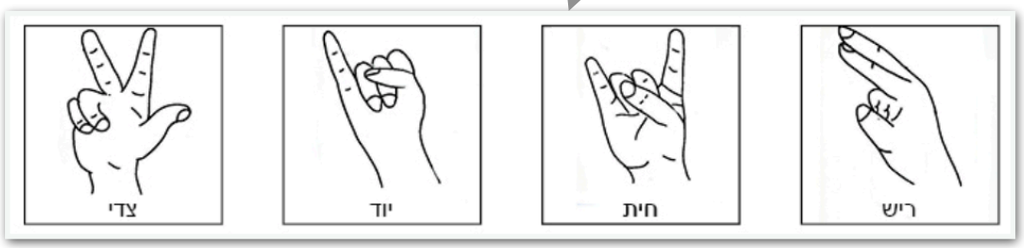
What is AAC? Augmentative and Alternative Communication
https://www.youtube.com/watch?v=r3m8_YmTDDM

Augmentative and Alternative Communication (AAC)

Unaided AAC

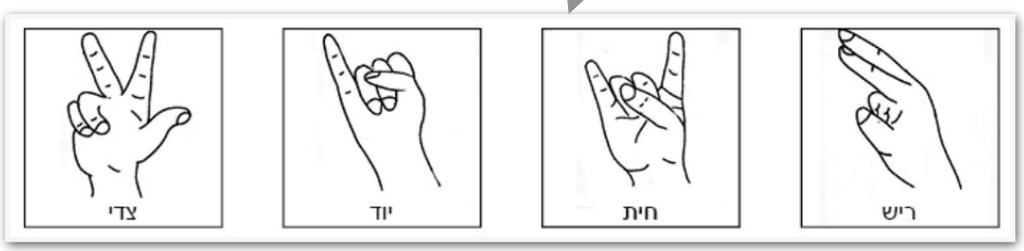
Augmentative and Alternative Communication (AAC)

Unaided AAC



Augmentative and Alternative Communication (AAC)

Unaided AAC



Augmentative and Alternative Communication (AAC)

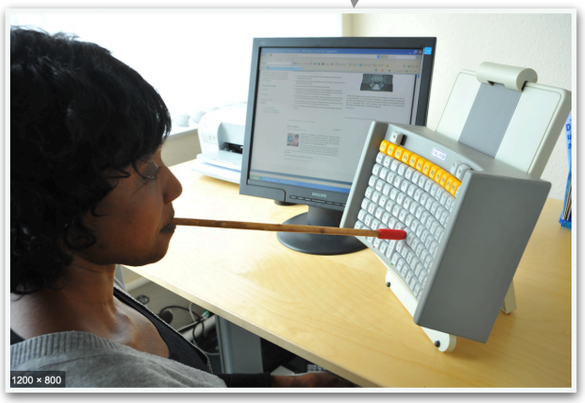
Unaided AAC

Augmentative and Alternative Communication (AAC)

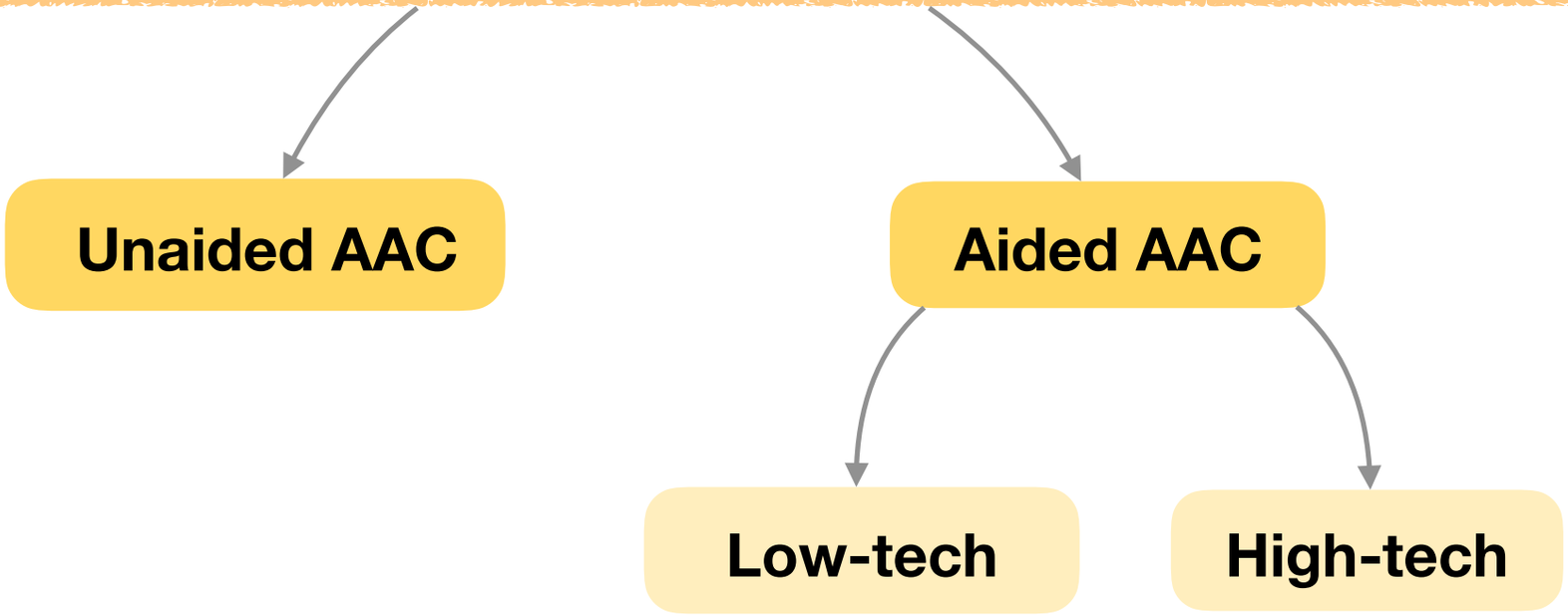
Unaided AAC

Aided AAC

Low-tech

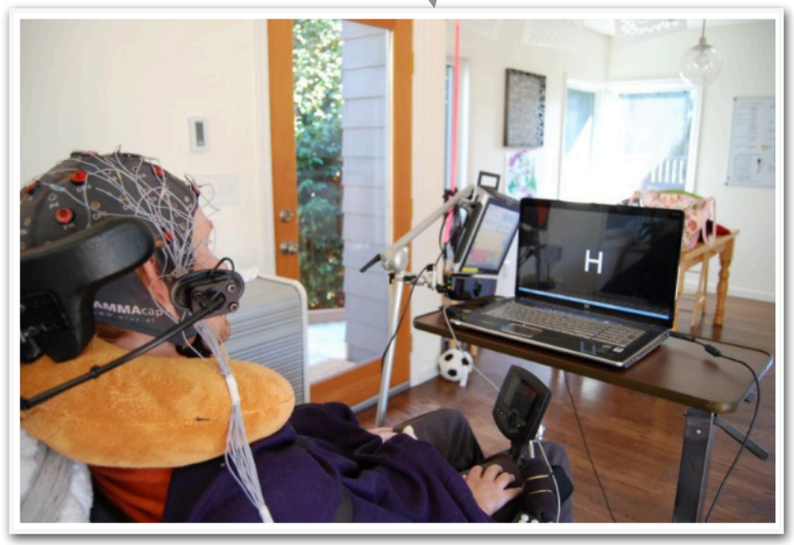


Augmentative and Alternative Communication (AAC)



Augmentative and Alternative Communication (AAC)

- **Unaided AAC**
- **Aided AAC**
  - **Low-tech**
  - **High-tech**



Augmentative and Alternative Communication (AAC)

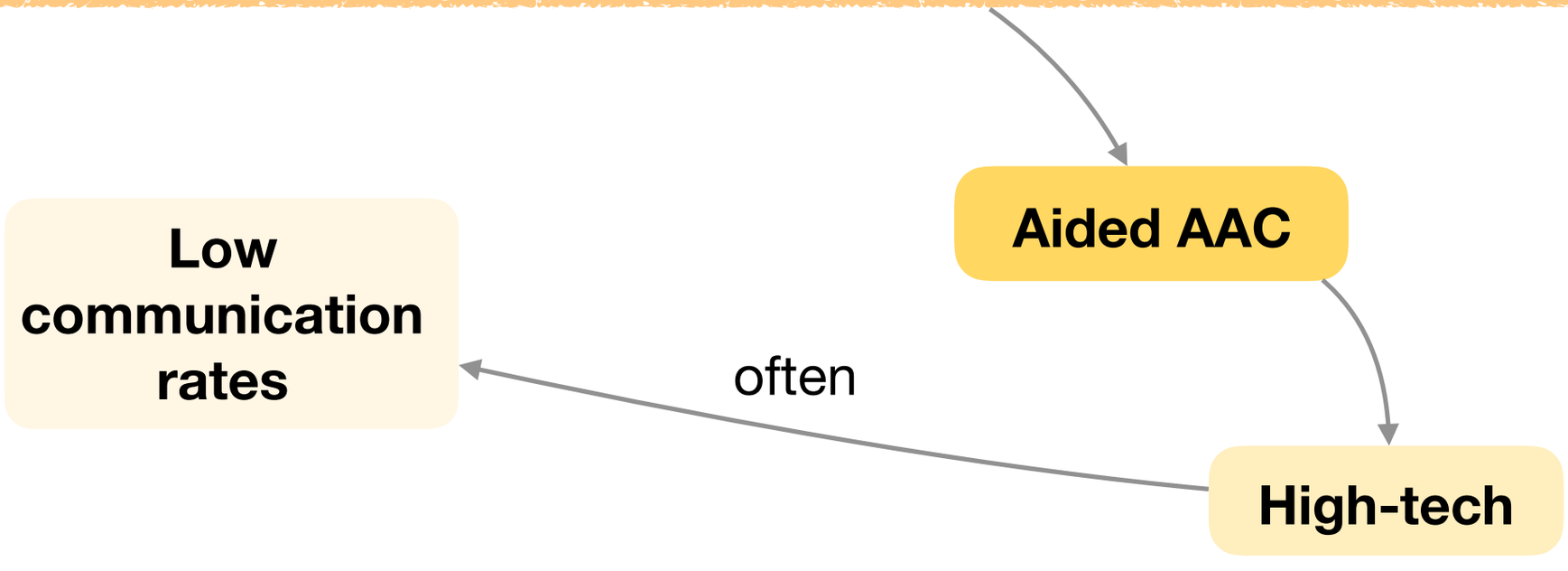
Unaided AAC

Aided AAC

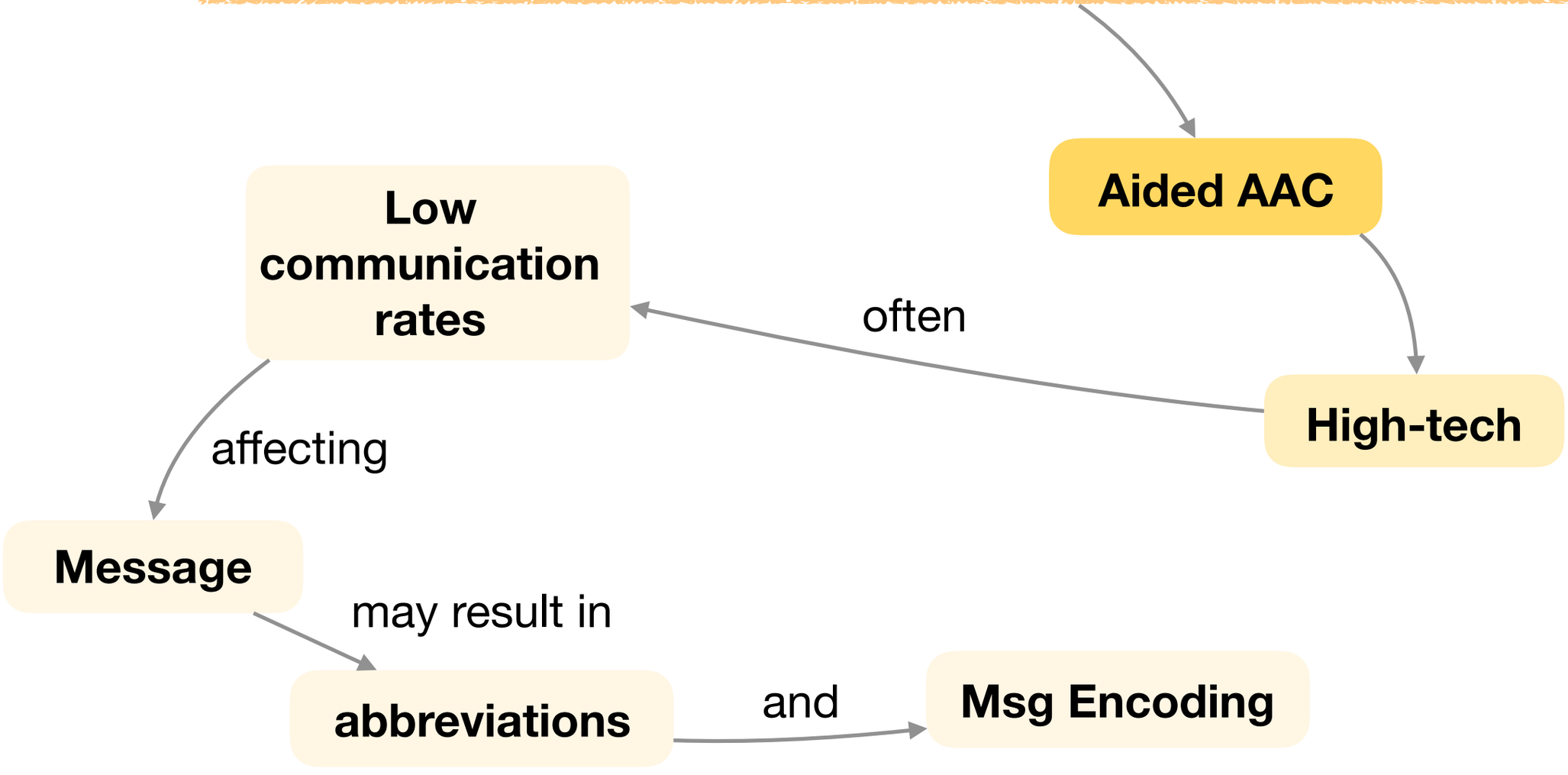
Low-tech

High-tech

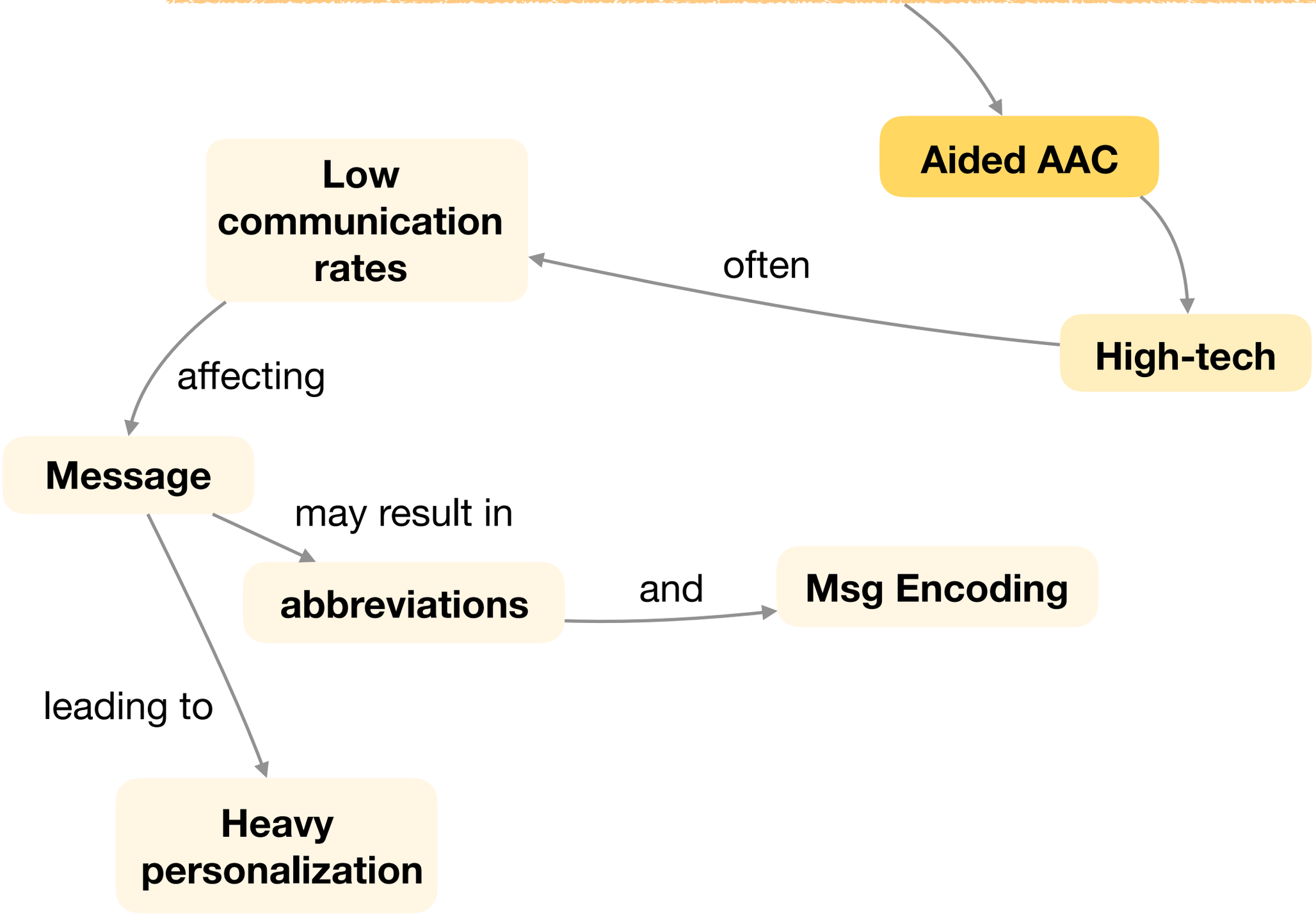
Augmentative and Alternative Communication (AAC)



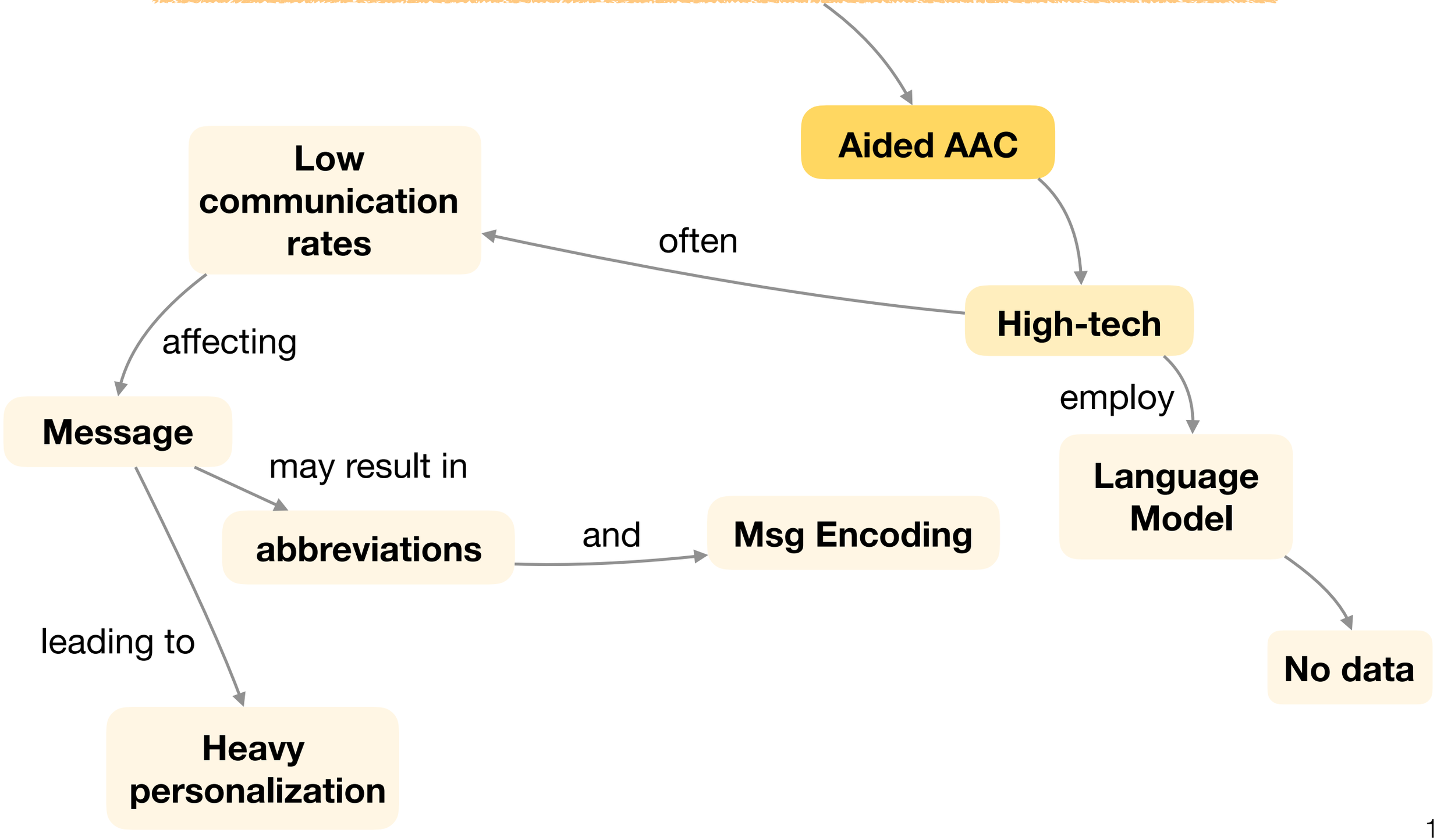
Augmentative and Alternative Communication (AAC)



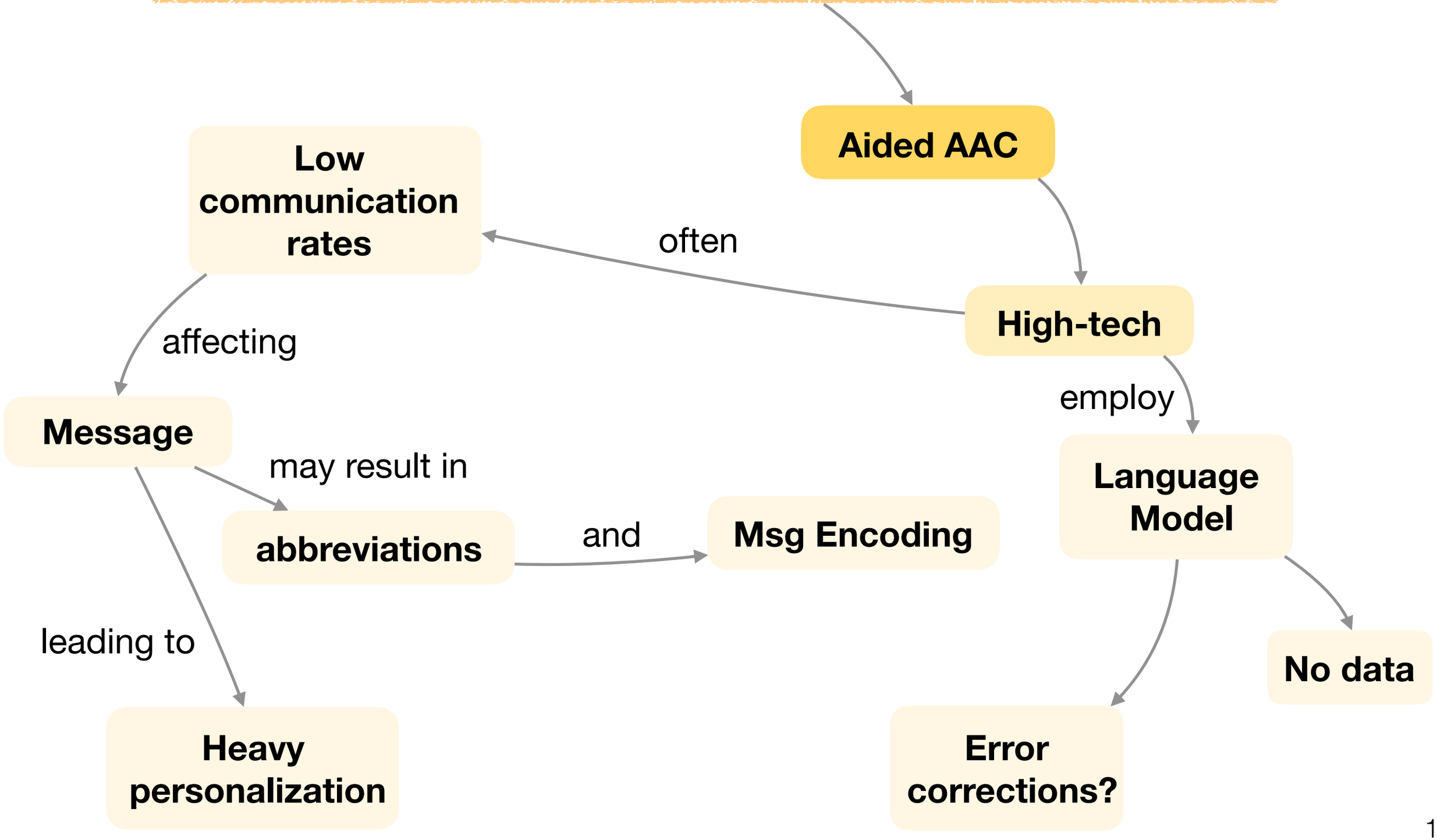
Augmentative and Alternative Communication (AAC)



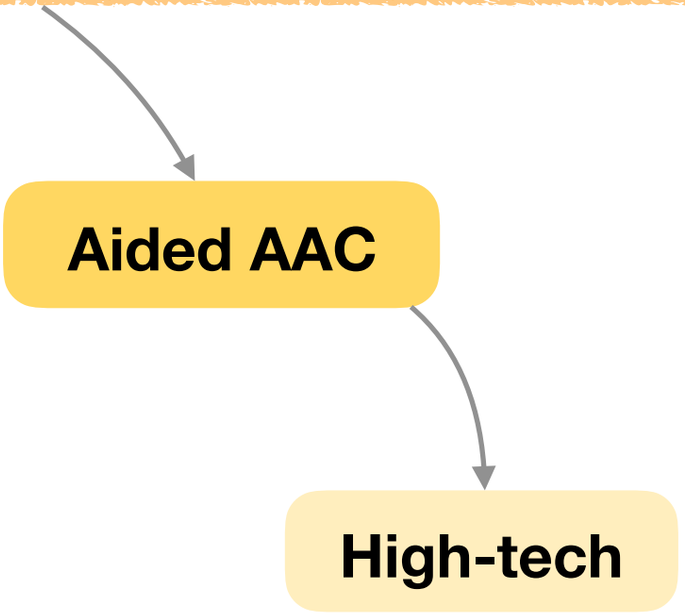
Augmentative and Alternative Communication (AAC)



Augmentative and Alternative Communication (AAC)



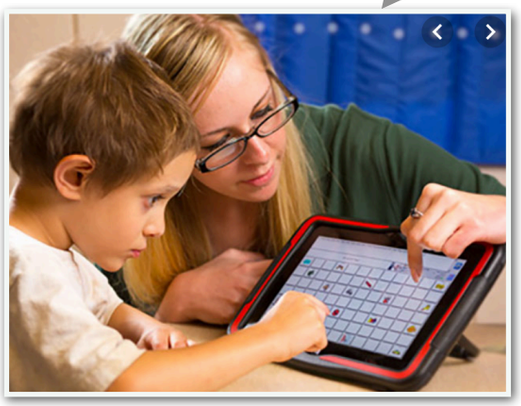
Augmentative and Alternative Communication (AAC)



Augmentative and Alternative Communication (AAC)

Aided AAC

High-tech



BCI

BCI

Brain-Computer Interface (BCI)

BCI

Brain-Computer Interface (BCI)

vary by

purpose

gaming

mobility

communication

BCI

Brain-Computer Interface (BCI)

vary by

technology

EEG

BCI

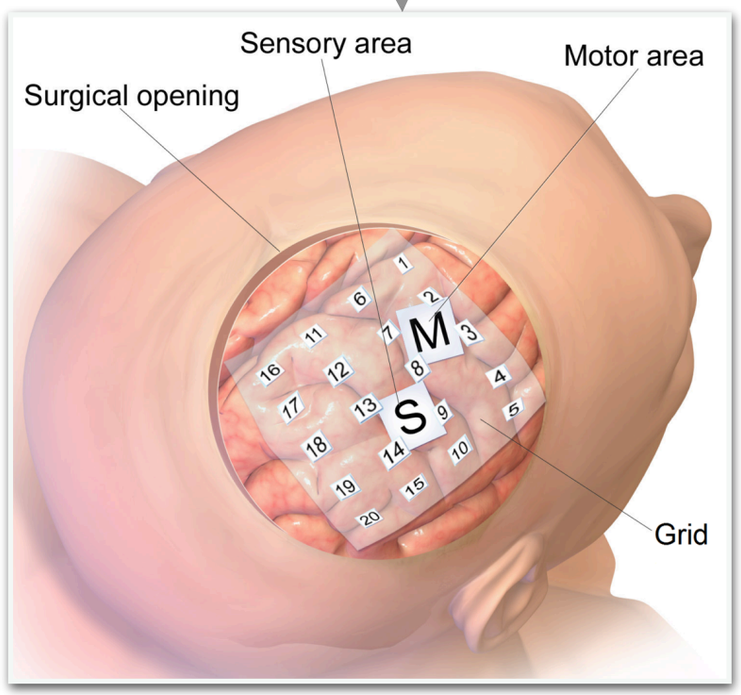
Brain-Computer Interface (BCI)

vary by

technology

TBD

EcoG



EEG

BCI

Brain-Computer Interface (BCI)

vary by

technology

EEG

BCI

Brain-Computer Interface (BCI)

vary by

technology

purpose

EEG

communication

BCI

Brain-Computer Interface (BCI)

vary by

technology

purpose

EEG

communication

elicit

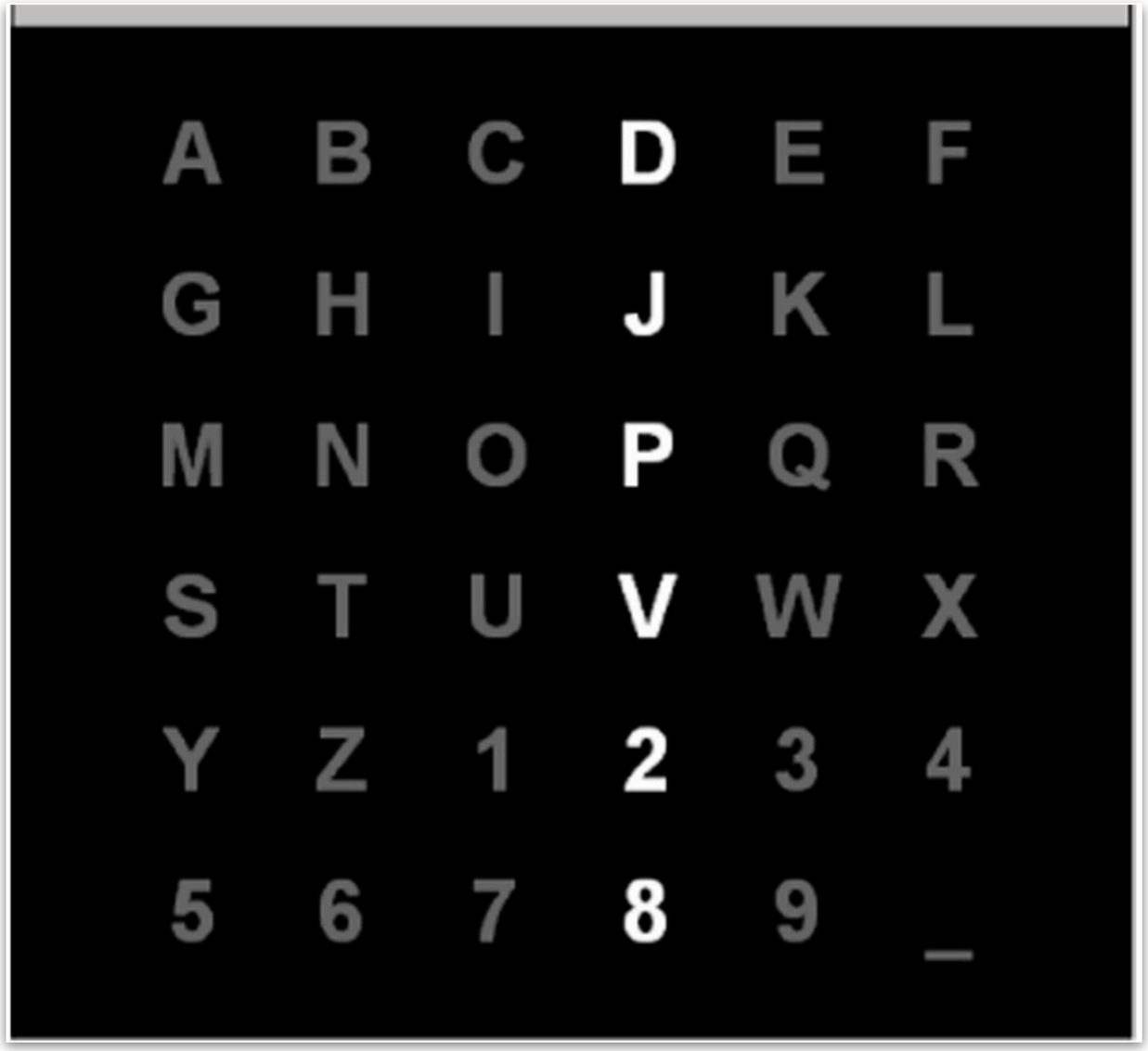
Oddball paradigm

P300

half second

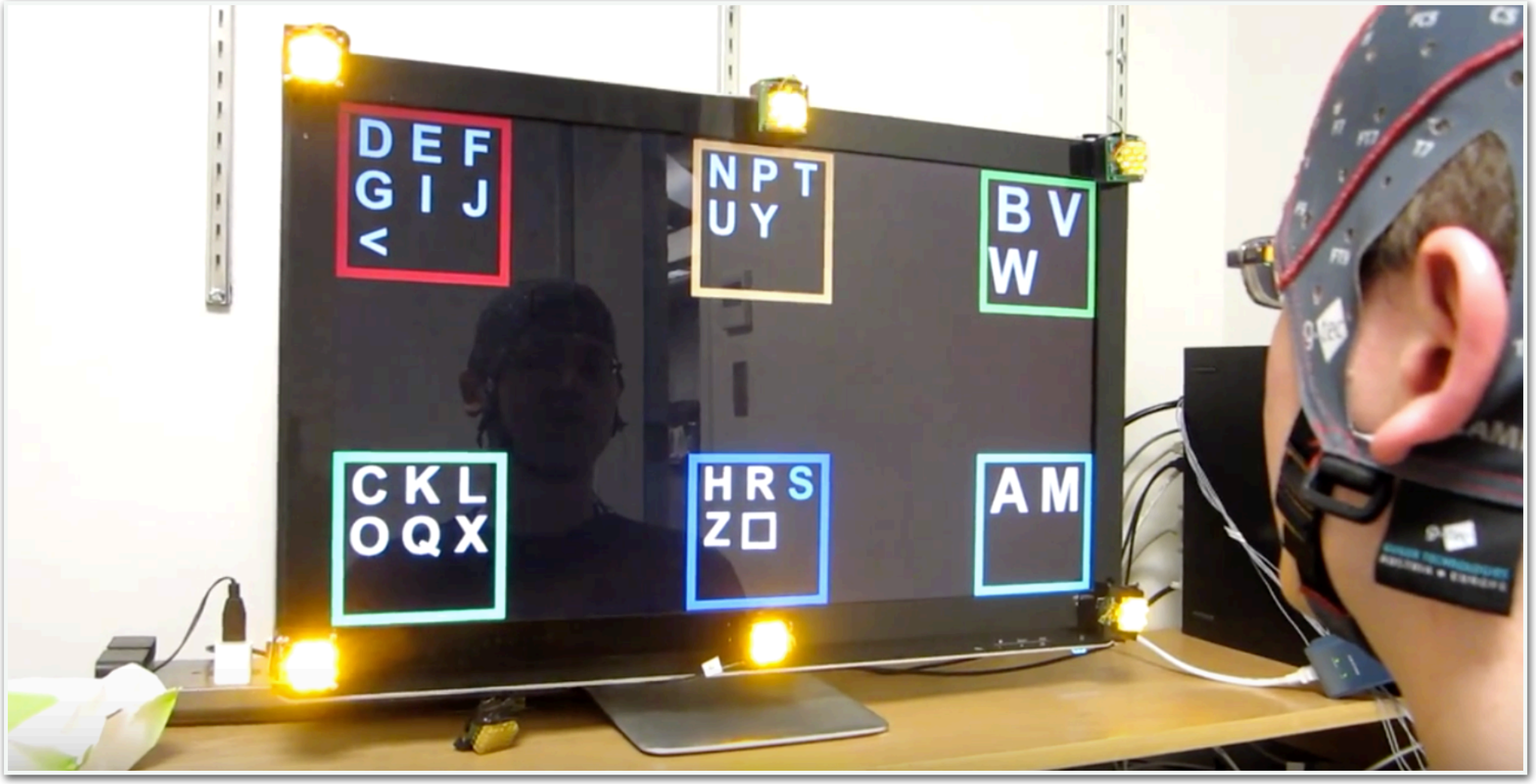
decision making

BCI



P300 speller

Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials, Farwell and Donchin, Electroencephalography and clinical Neurophysiology, 1988



Steady State Visual Evoked Potentials (SSVEP shuffle speller)

Spelling with Steady-State Visual Evoked Potentials, Friman et al. , EMBS Conference on Neural Engineering, 2007

Rapid Serial Visual Presentation (RSVP) Keyboard

RSVP Keyboard: An EEG Based Typing Interface, Orhan et al. Proc IEEE Acoust Speech Signal Processing, 2012

Typical text entry prediction

direct selection

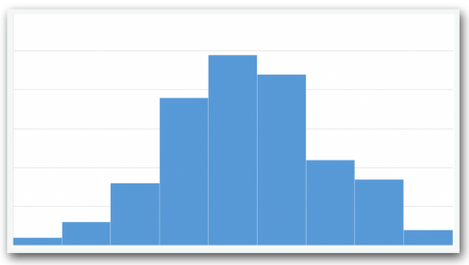
Fast

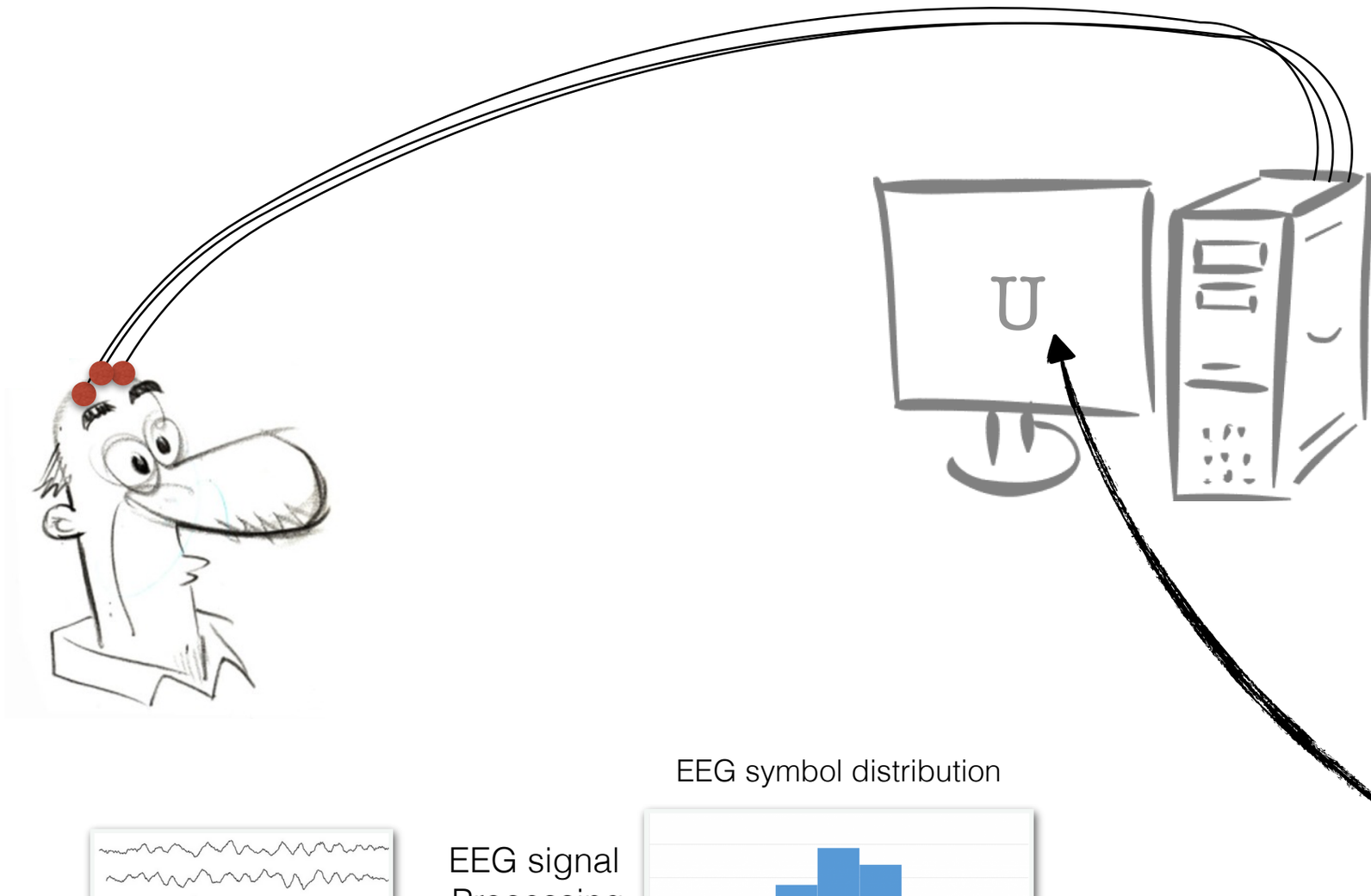
RSVP keyboard

binary response for symbols

Slow

symbol selection probability distribution




U
EEG signal
Processing
EEG symbol distribution
Candidate symbols

Naive approach:

Signal acquisition and processing for each of the symbols in the system

Naive approach:

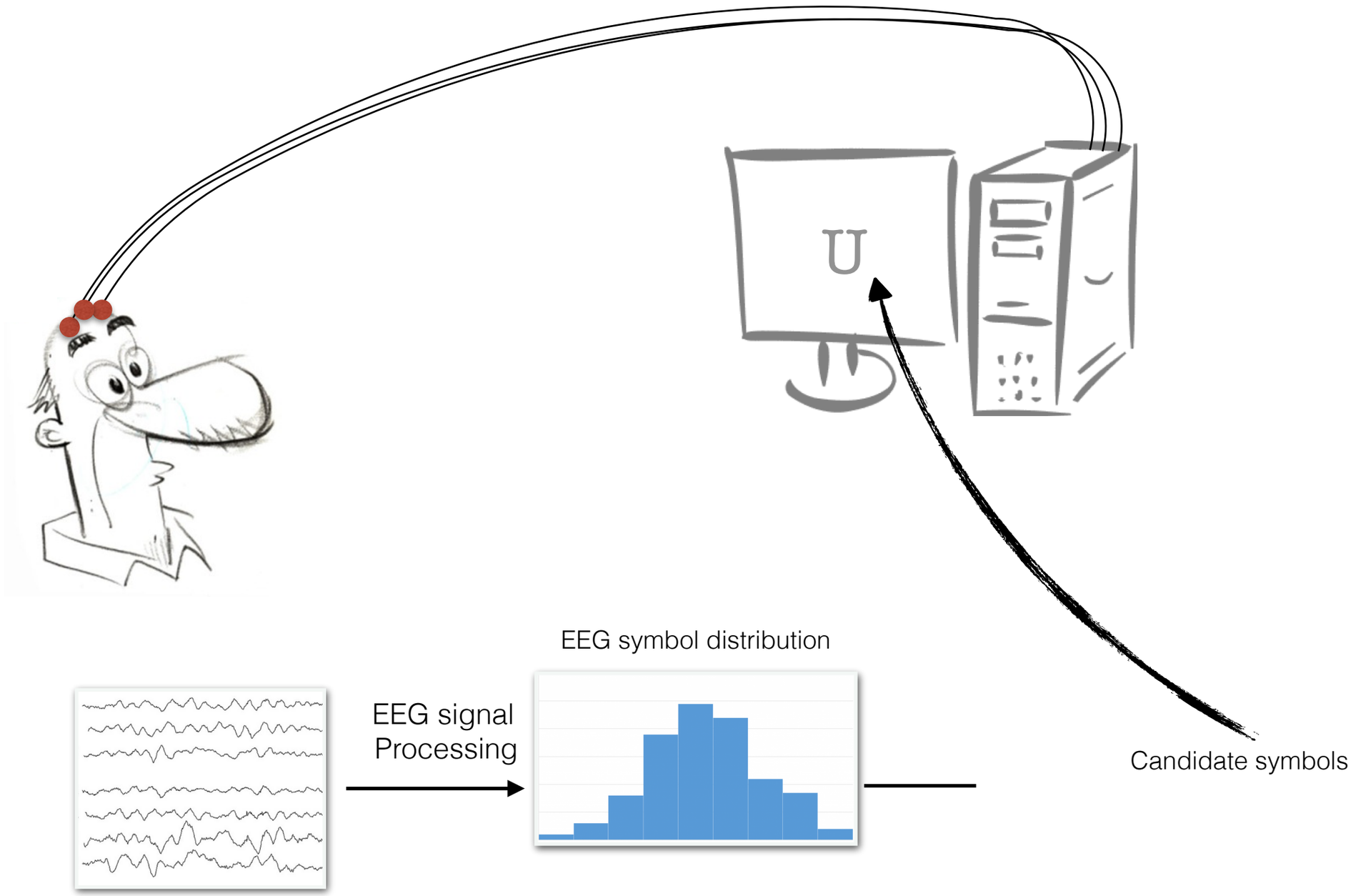
Signal acquisition and processing for each of the symbols in the system

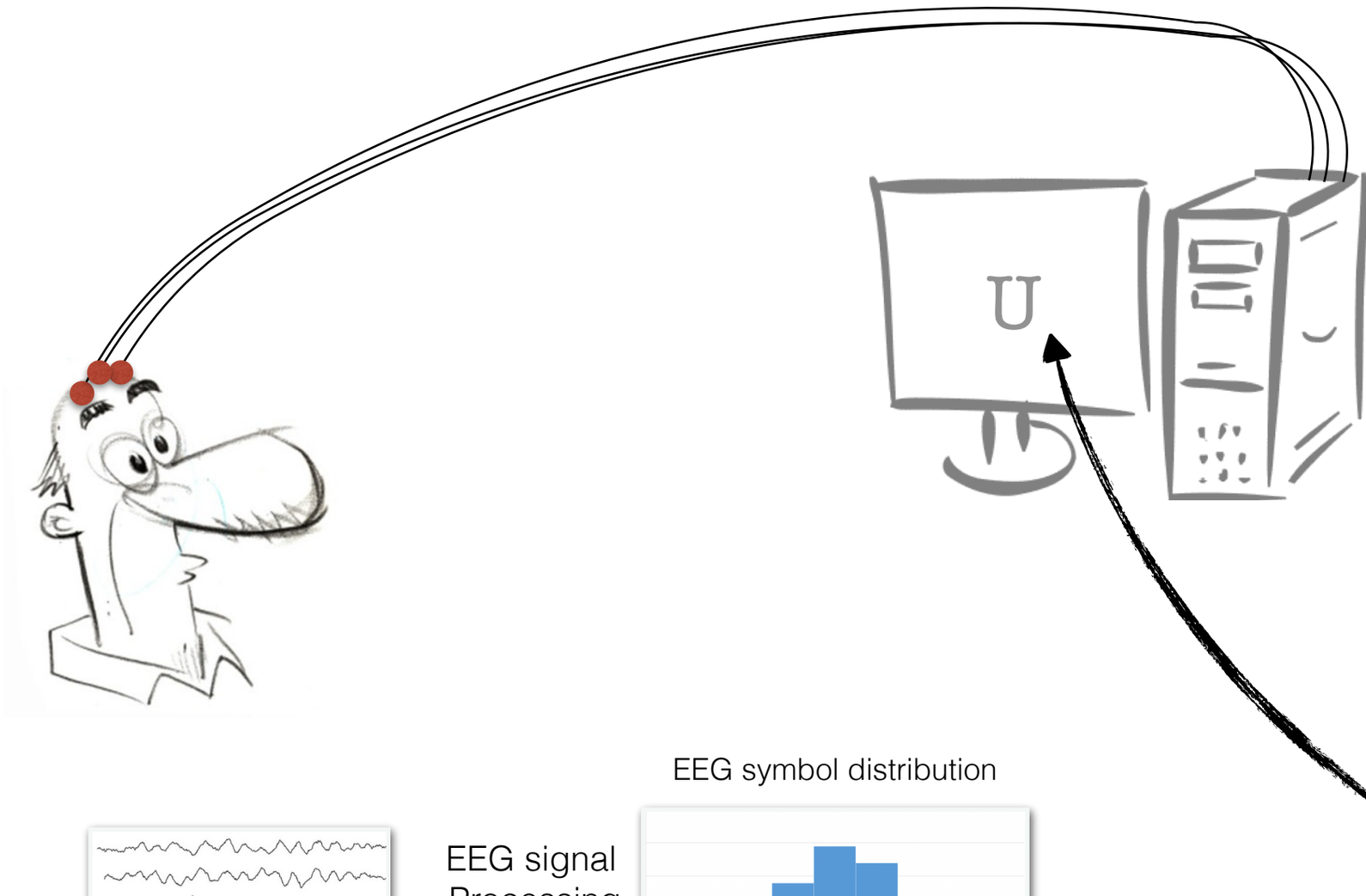
Naive approach:

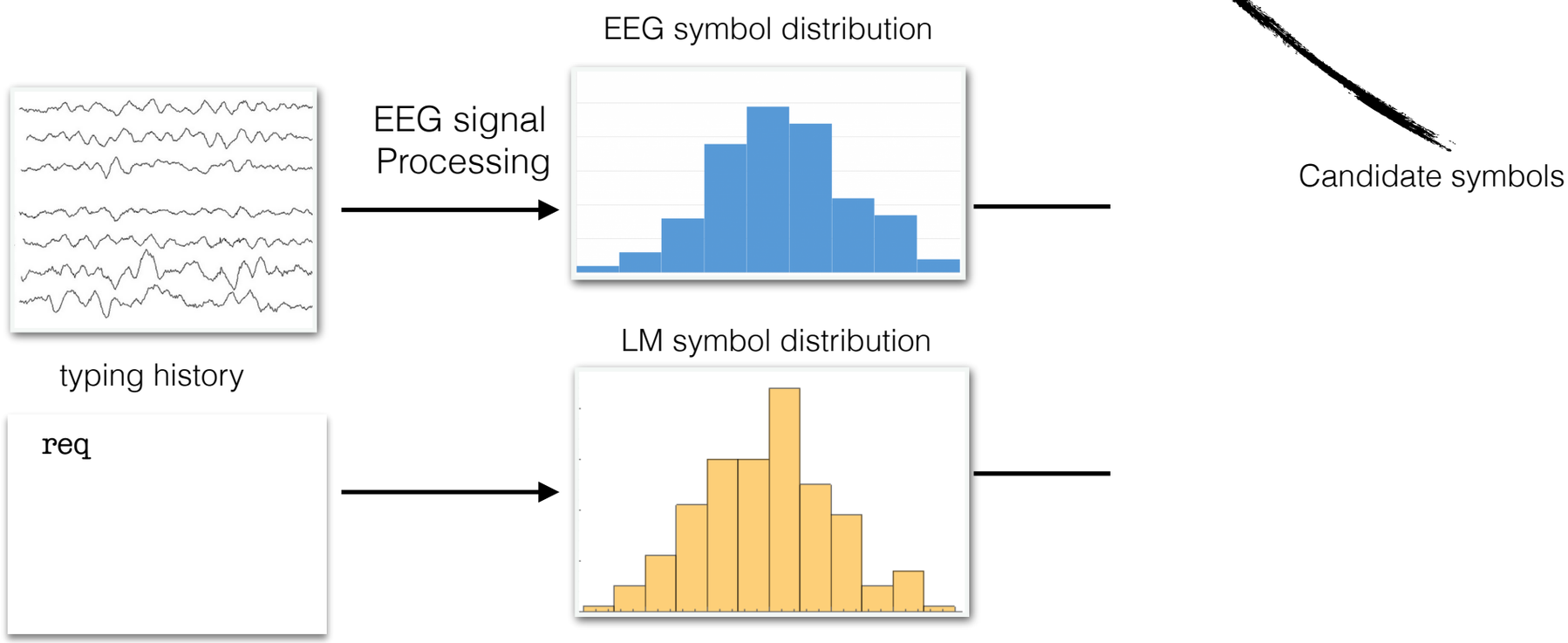
Signal acquisition and processing for each of the symbols in the system

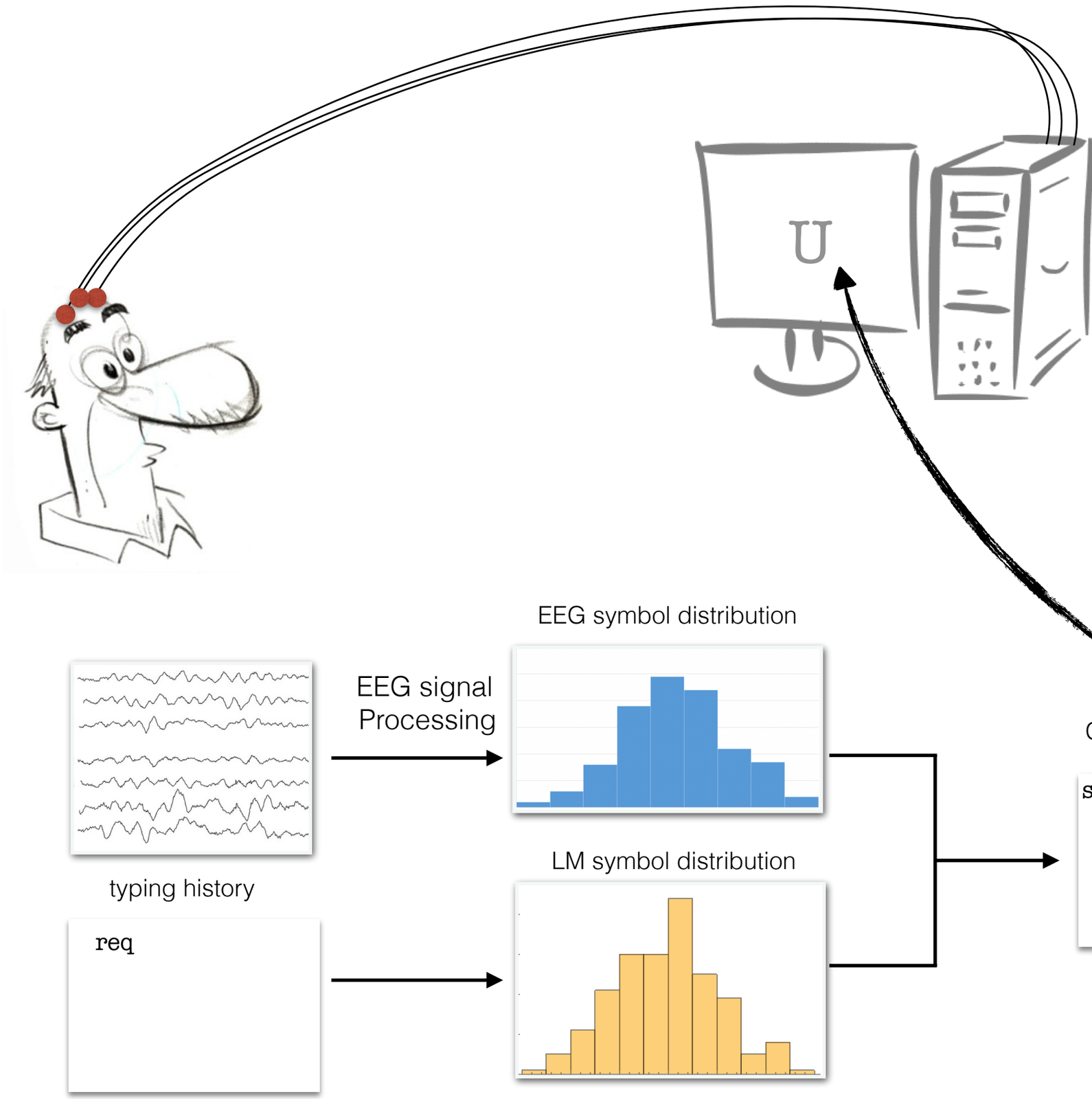
Language Model (LM) approach:

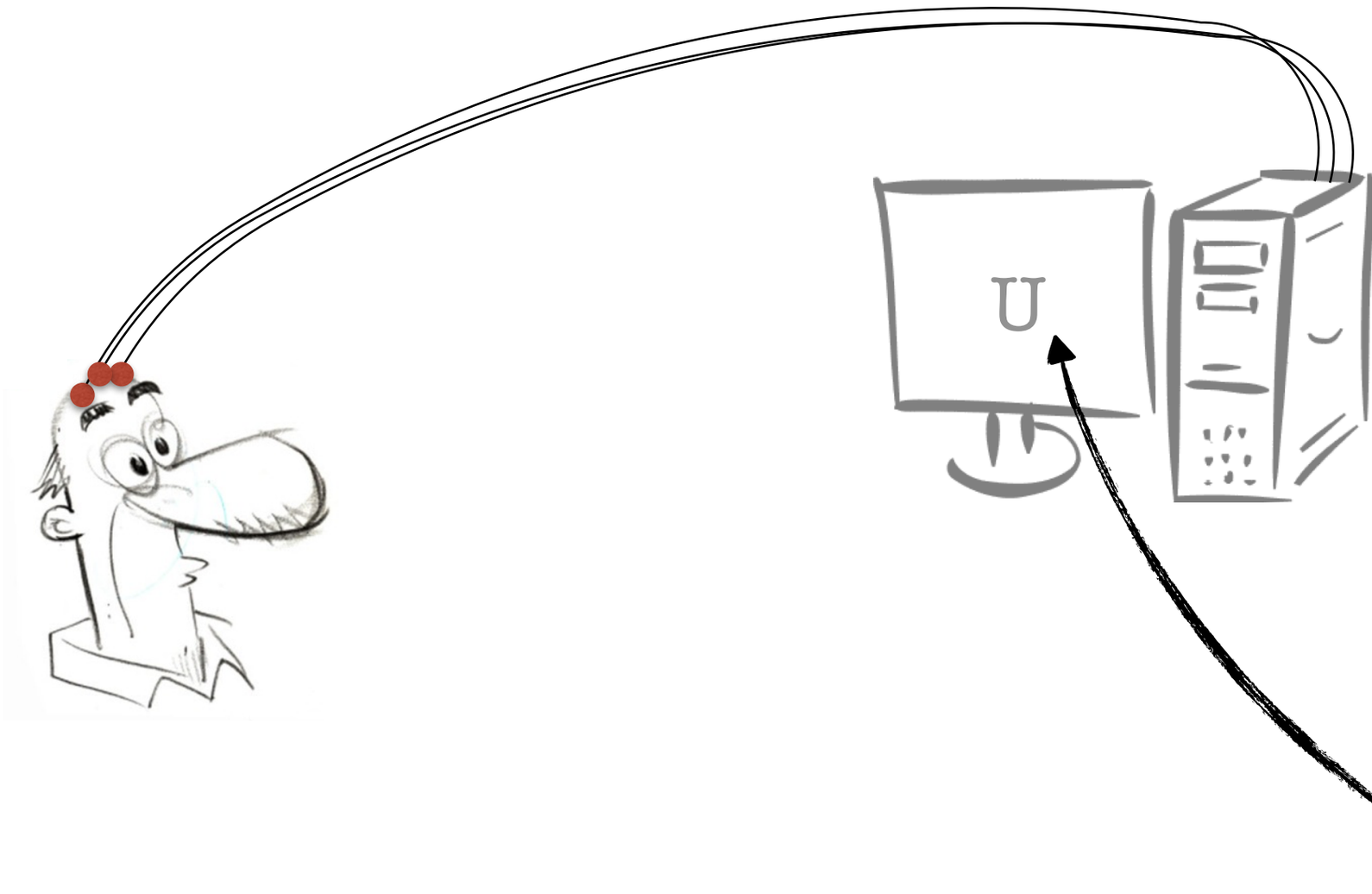
Signal acquisition and processing for fewer symbols than overall symbols using language patterns


U
EEG signal
Processing
EEG symbol distribution
Candidate symbols

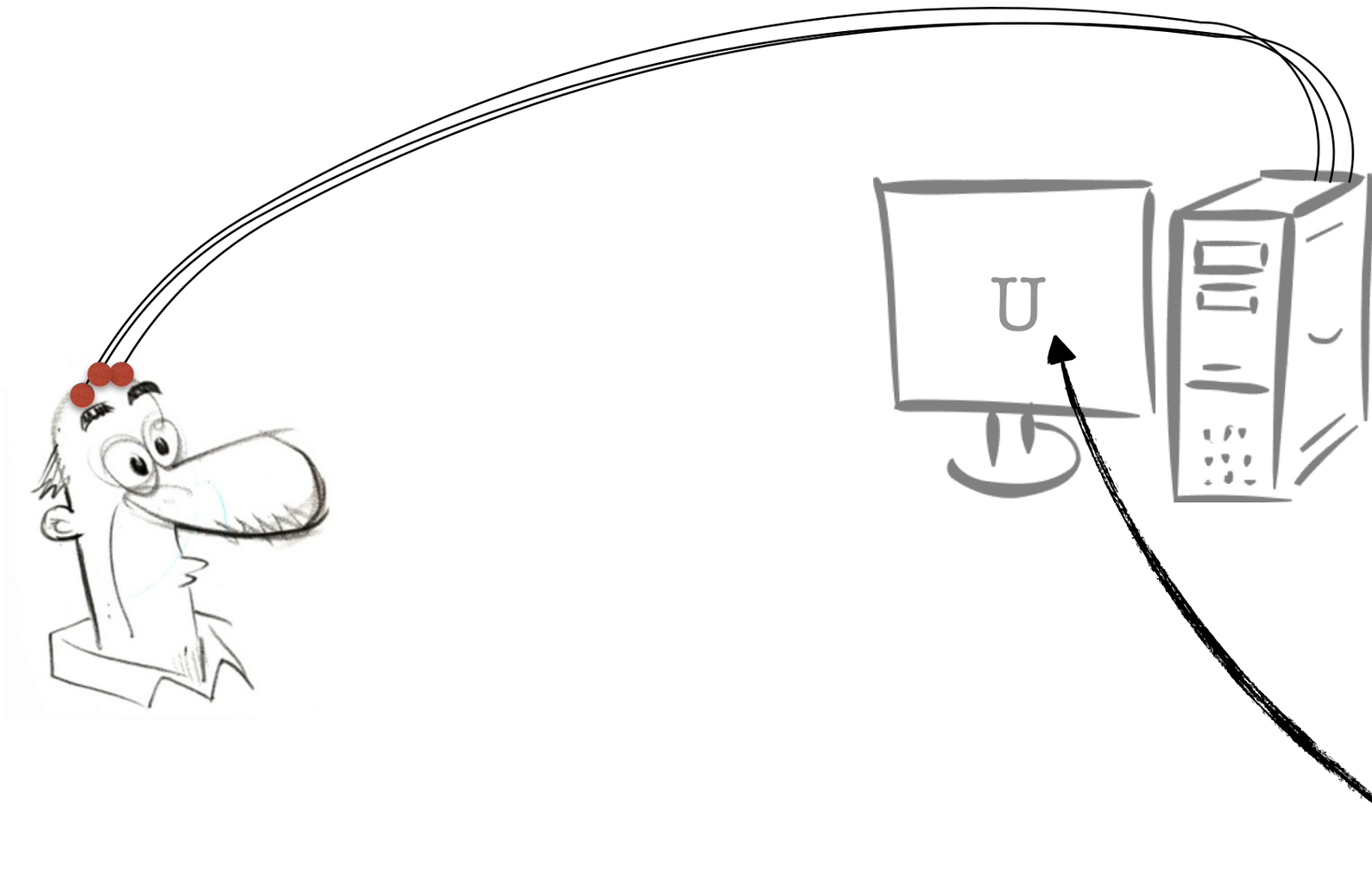

U
Candidate symbols


EEG symbol distribution
EEG signal
Processing
typing history
req
LM symbol distribution


U
EEG symbol distribution
EEG signal
Processing
typing history
req
LM symbol distribution
Candidate symbols
symbol: confidence
u: 0.95
m: 0.05



$$P(sym|EEG) = \frac{P(sym)P(EEG|sym)}{P(EEG)}$$



$$P(sym|EEG) \propto P(sym)P(EEG|sym)$$

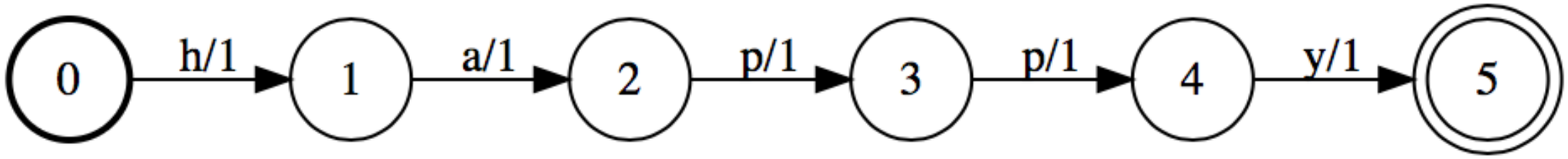
symbol: confidence

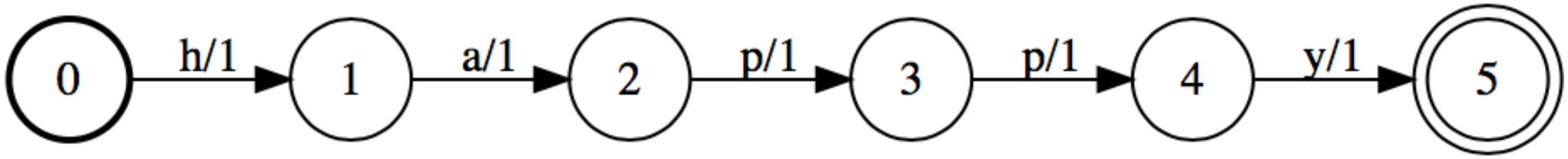
u: 0.95
m: 0.05

A Toy Example

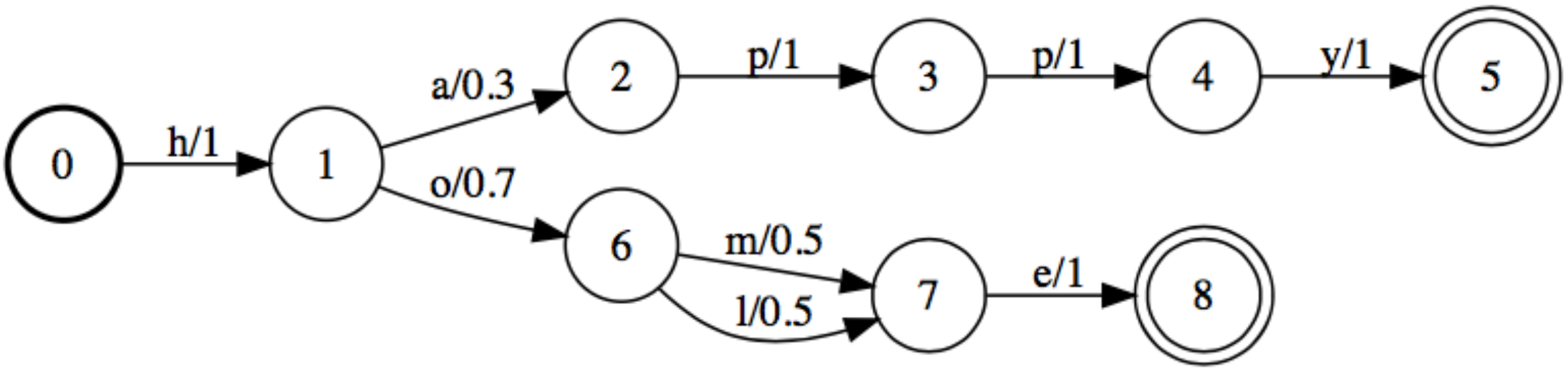
vocabulary:
happy
home
hole

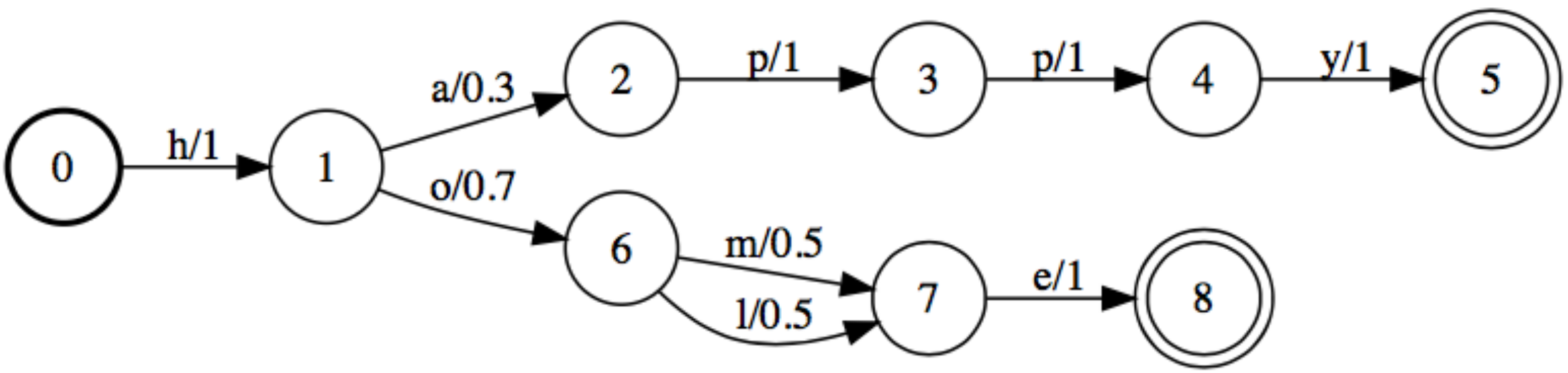
vocabulary:
~~happy~~
home
hole




0
h/1
1
a/1
2
p/1
3
p/1
4
y/1
5

vocabulary:
~~happy~~
~~home~~
~~hole~~



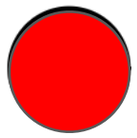


Compact representation of prob. dist. over set of strings
Defined by states, arcs, symbols, and weights

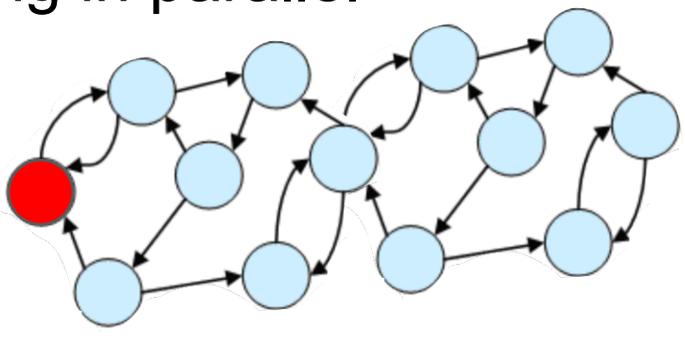
A Basic Language Model

Basic LM Module implementation

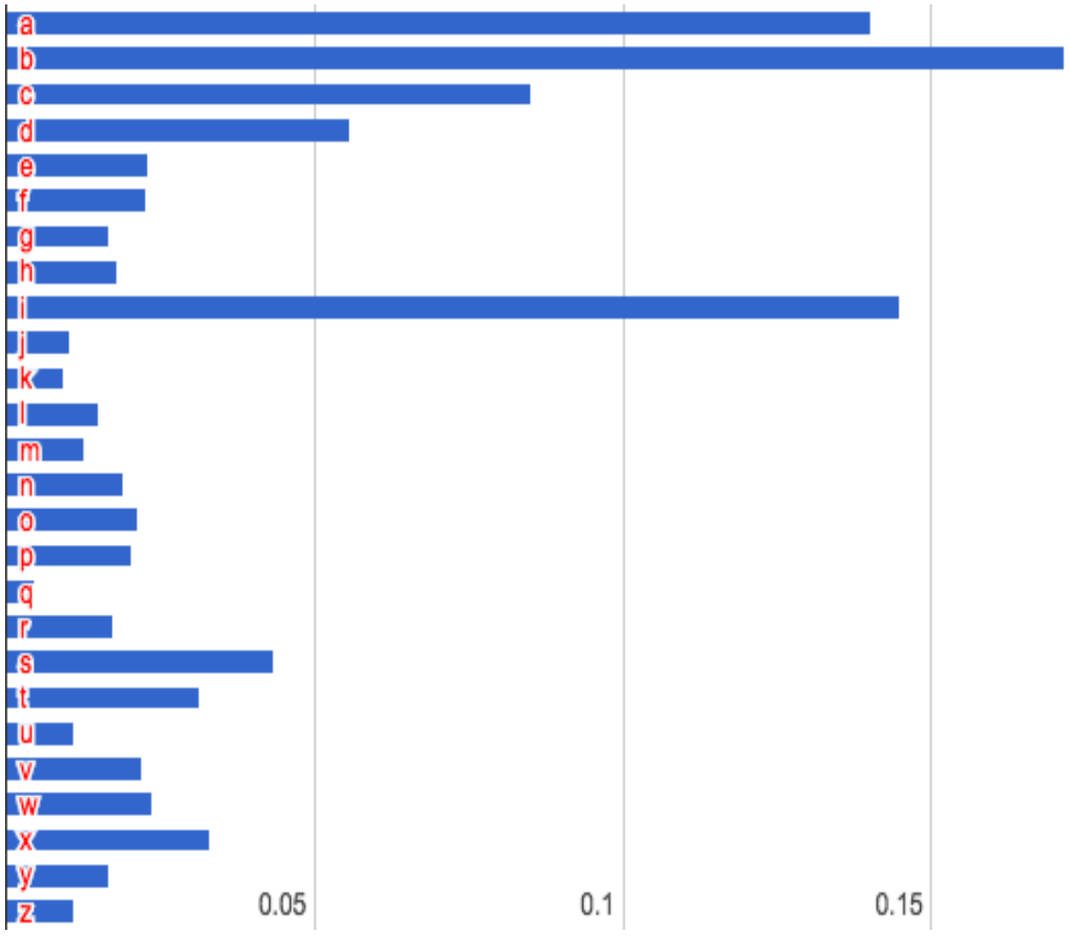
History



Intersect history and LM by walking in parallel



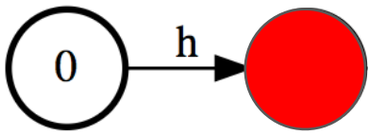
Language Model



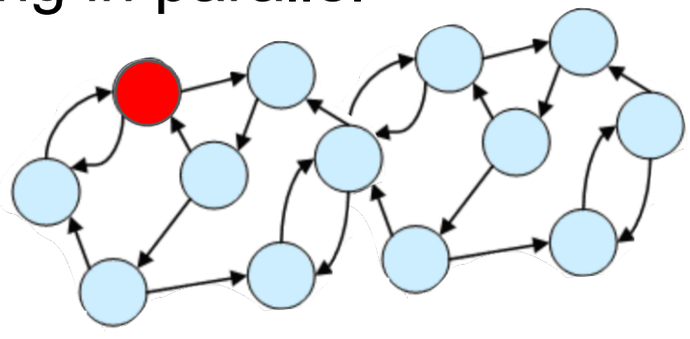
Probability of next character

Basic LM Module implementation

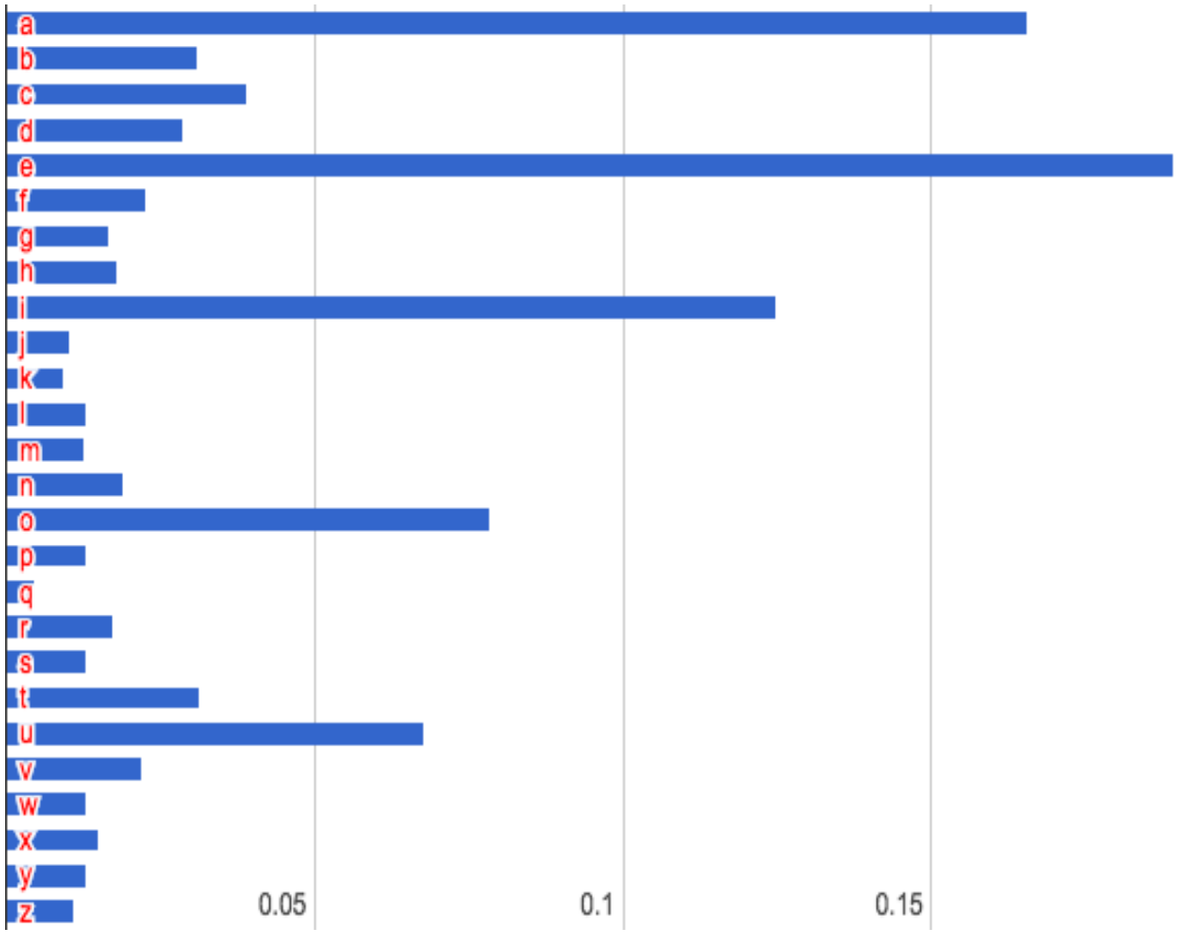
History



Intersect history and LM by walking in parallel



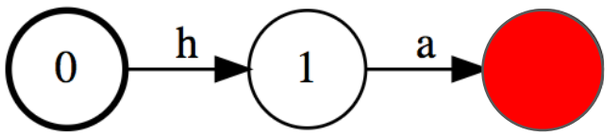
Language Model



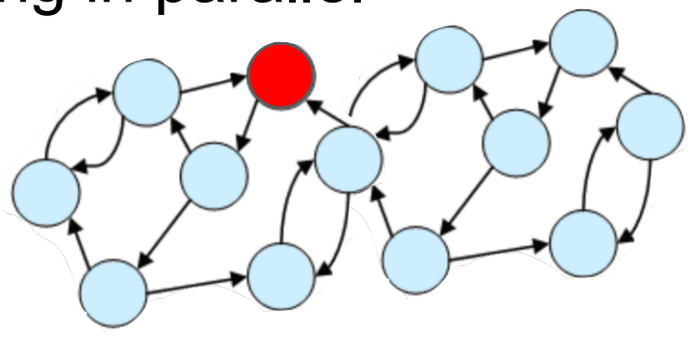
Probability of next character

Basic LM Module implementation

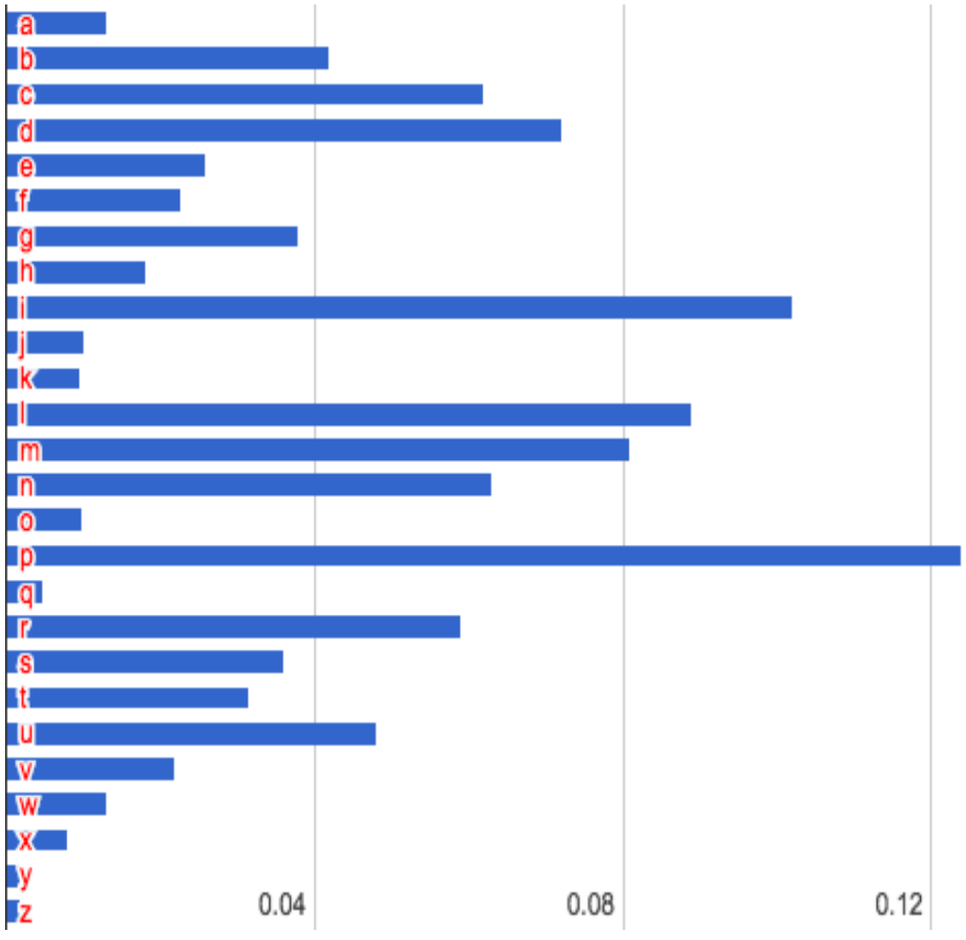
History



Intersect history and LM by walking in parallel



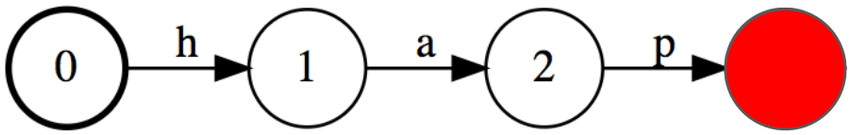
Language Model



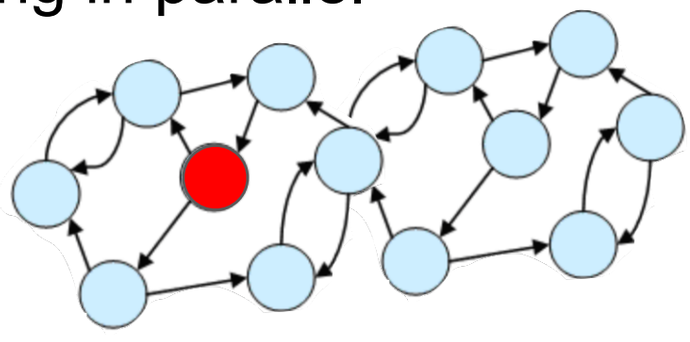
Probability of next character

Basic LM Module implementation

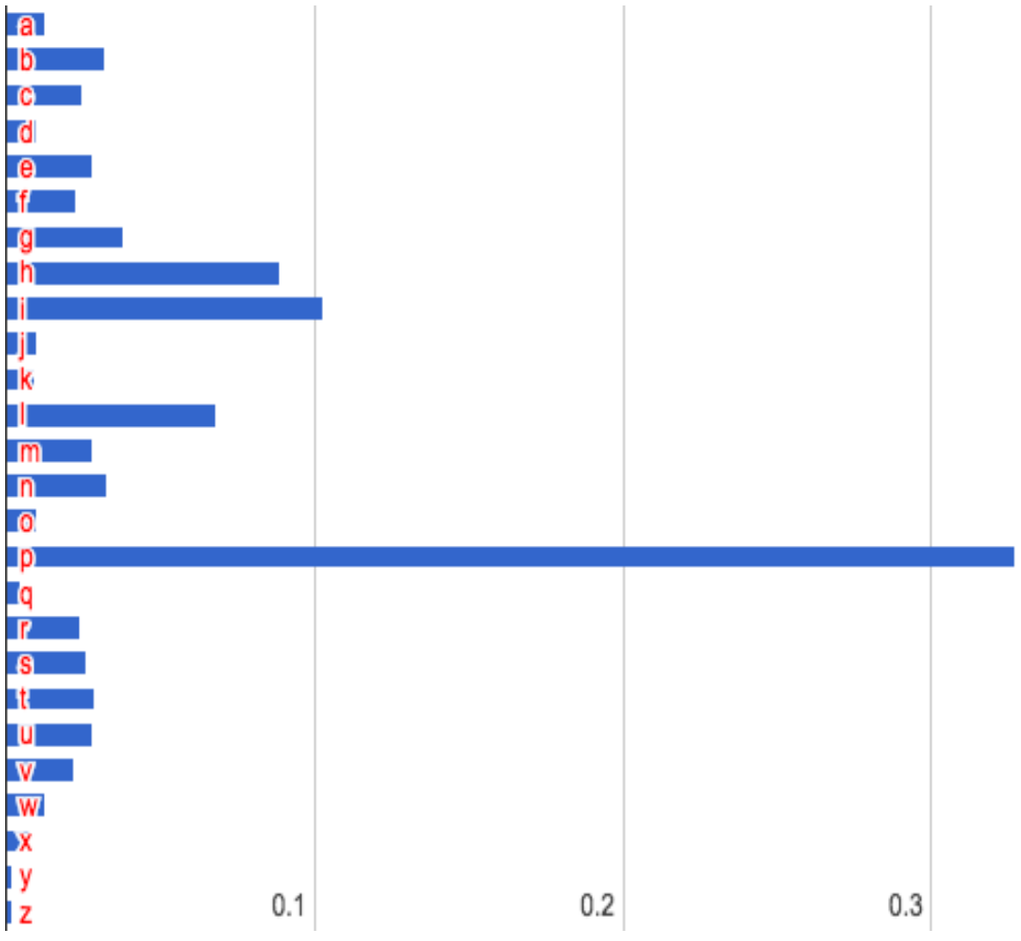
History



Intersect history and LM by walking in parallel

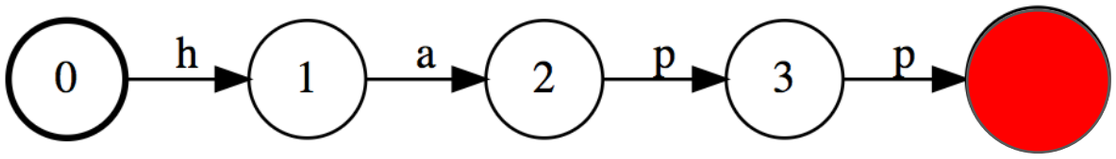


Language Model

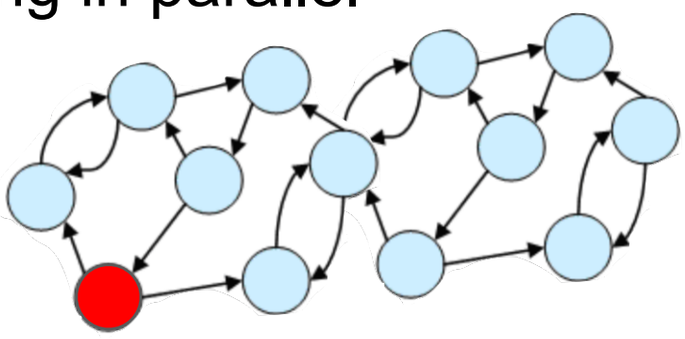


Probability of next character

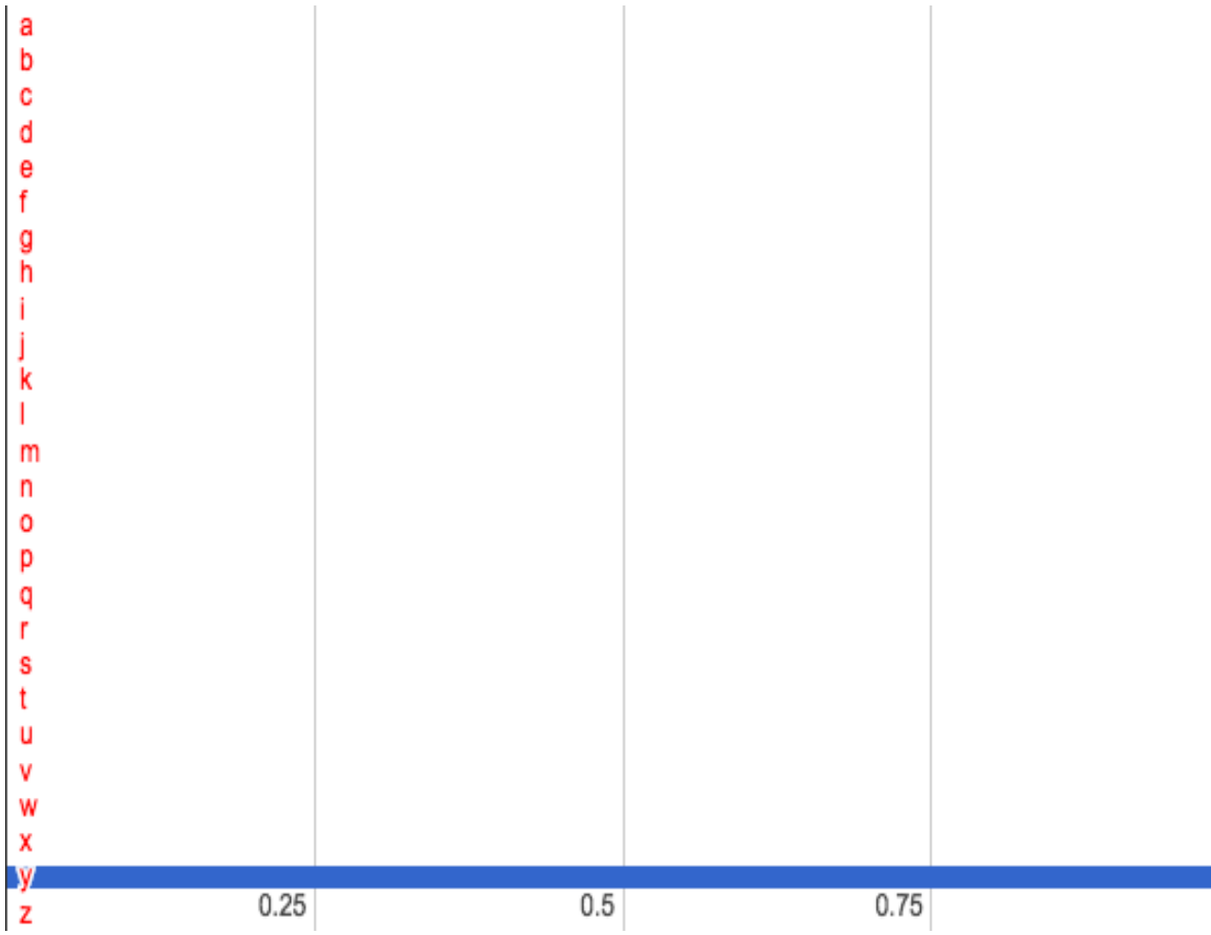
Basic LM Module implementation



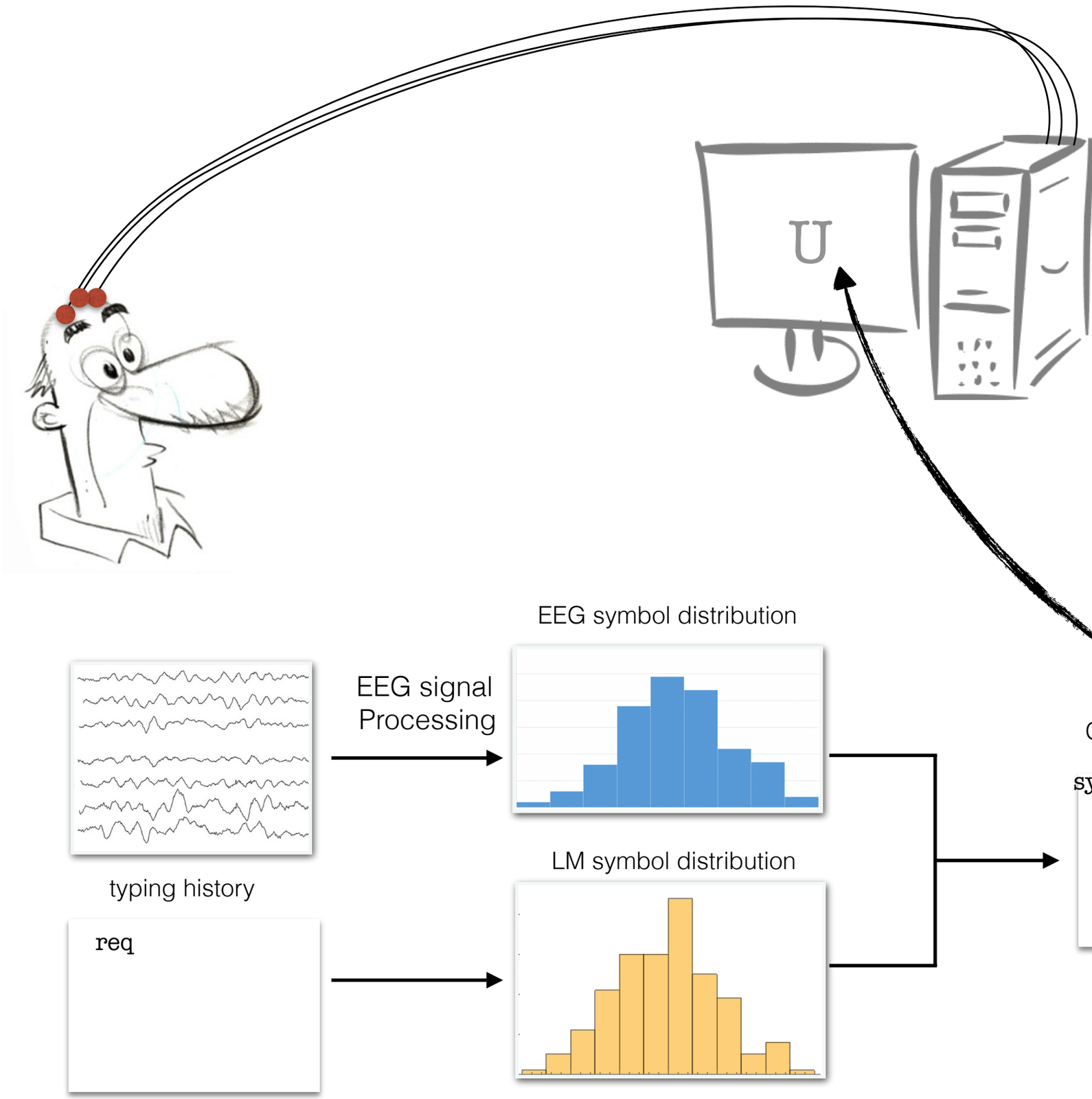
Intersect history and LM by walking in parallel



Language Model



Probability of next character


U
EEG symbol distribution
EEG signal
Processing
typing history
req
LM symbol distribution
Candidate symbols
symbol: confidence
u: 0.95
m: 0.05

$$P(sym|EEG) \propto P(sym)P(EEG|sym)$$

$$P(sym) = P(sym|X_m)$$
$$X_0 = \{sym_{t-i}\}, i > 0$$
$$X_1 = \{word_{t-i}\}, i > 0$$
$$X_2 = \{EEG_{t-i}\}, i \geq 0$$

Candidate symbols

symbol: confidence

u: 0.95
m: 0.05

$$P(sym|EEG) \propto P(sym)P(EEG|sym)$$

$$P(sym) = P(sym|X_m)$$

✓ $X_0 = \{sym_{t-i}\}, i > 0$

$X_1 = \{word_{t-i}\}, i > 0$

$X_2 = \{EEG_{t-i}\}, i \geq 0$

Candidate symbols

symbol: confidence

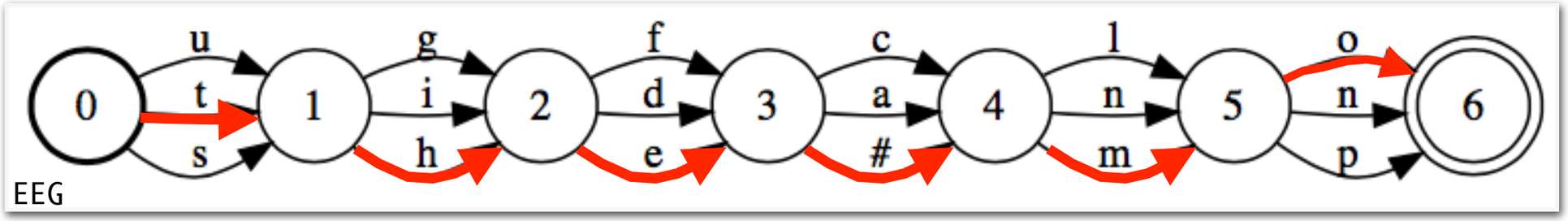
u: 0.95
m: 0.05

Issues with the basic LM module

In reality the EEGs are a distribution of symbols. While the basic model assumes a deterministic history.

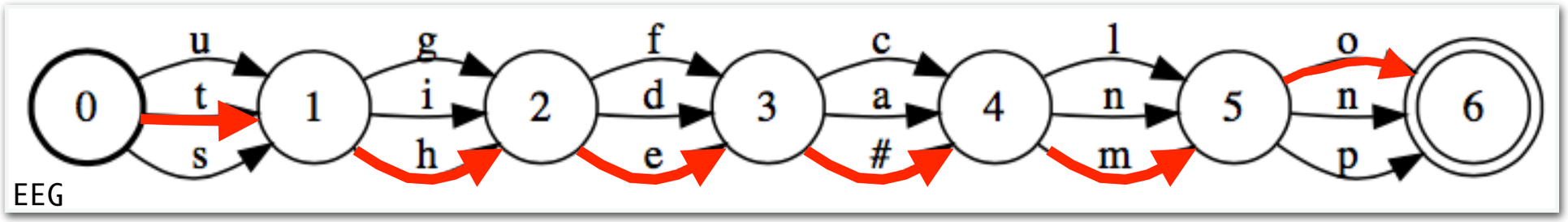
While the basic model learns probabilities from n-grams, it does not incorporate word level information.

The OCLM Language Model

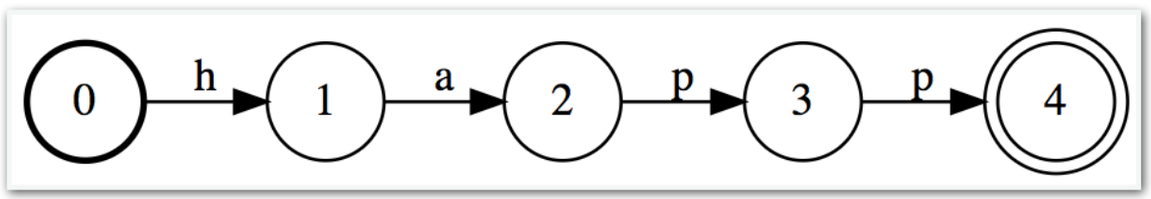


EEG

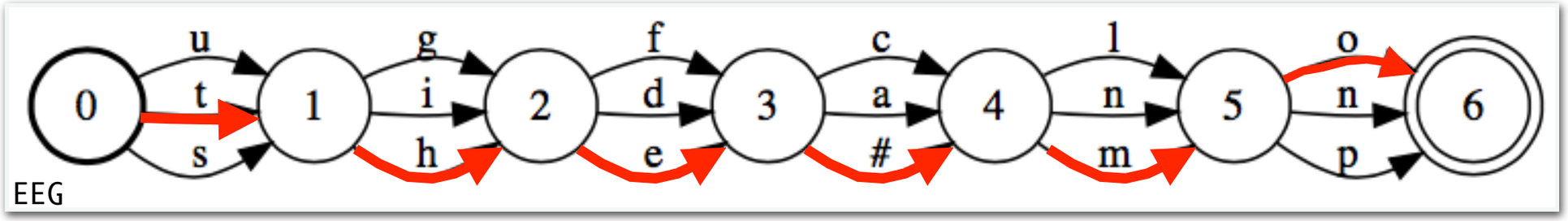
non-determinstic (n=3)

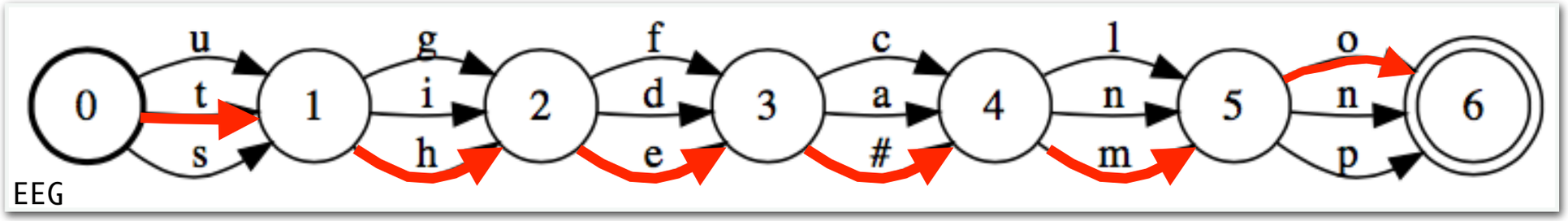


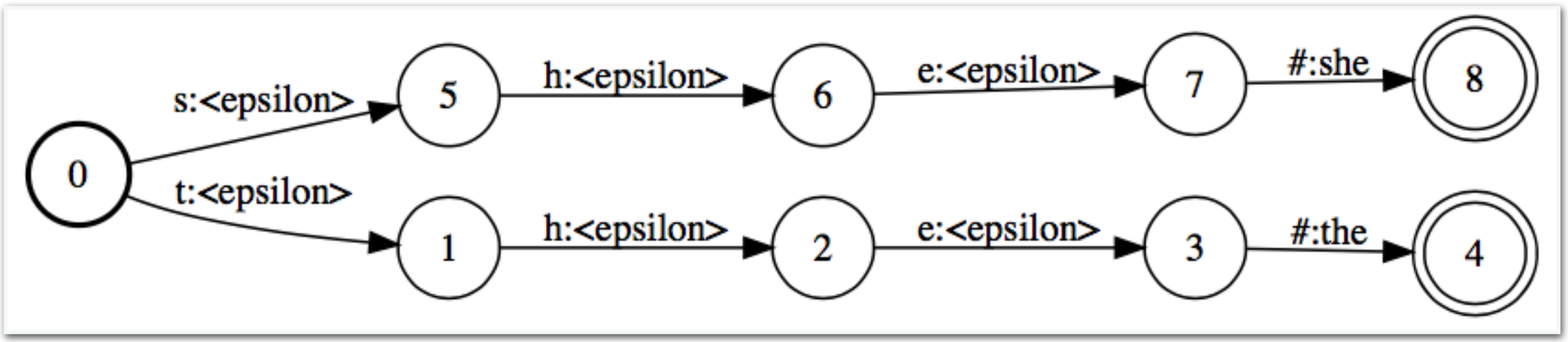
Determinstic (n=1)

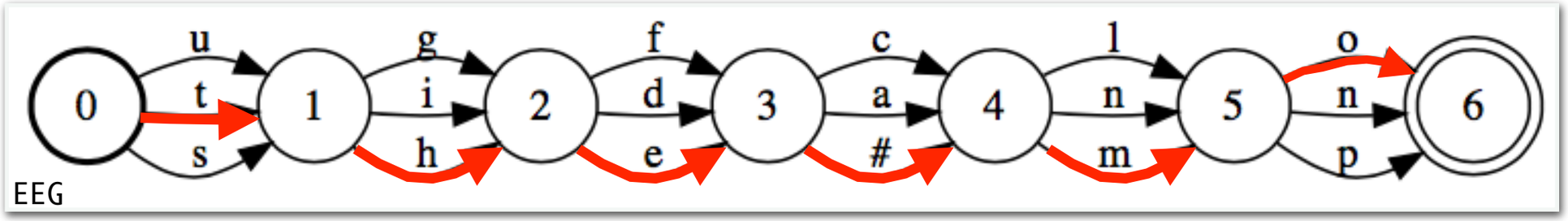


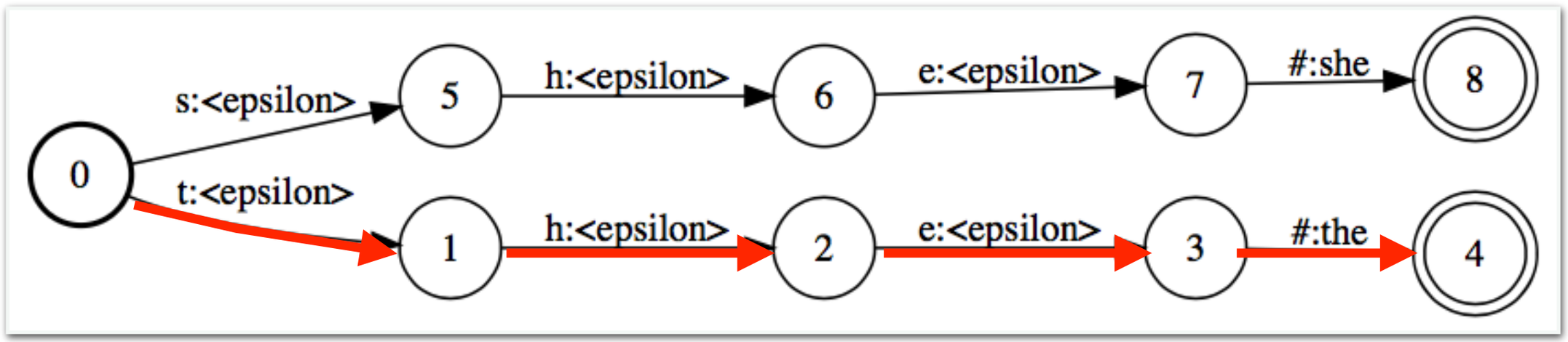
non-determinstic (n=3)

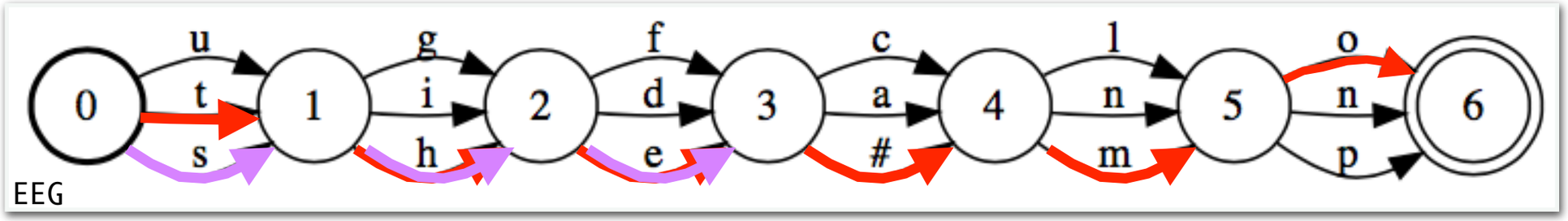


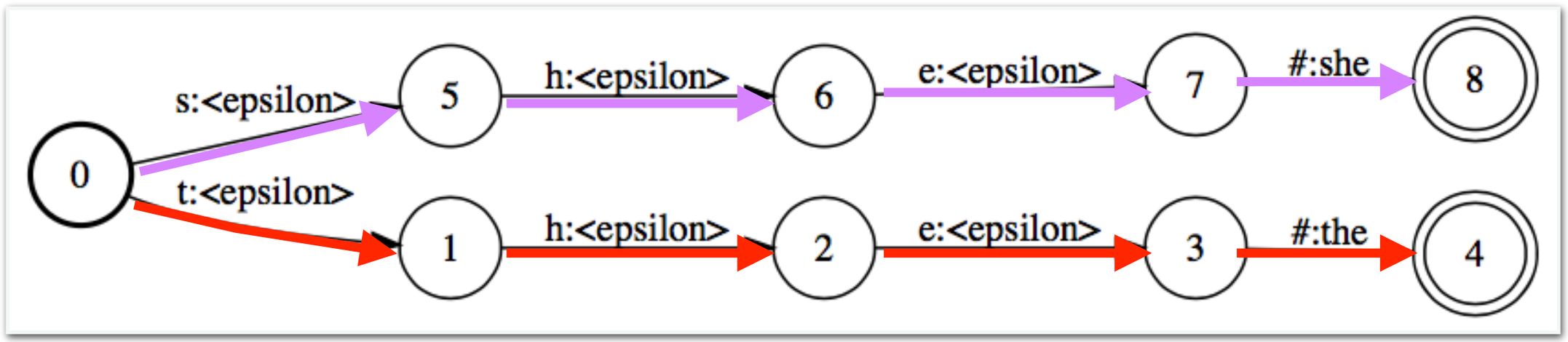

u
t
s
g
i
h
f
d
e
c
a
#
l
n
m
o
n
p
0
1
2
3
4
5
6
EEG


s:<epsilon>
t:<epsilon>
h:<epsilon>
h:<epsilon>
e:<epsilon>
e:<epsilon>
#:she
#:the
0
1
2
3
4
5
6
7
8

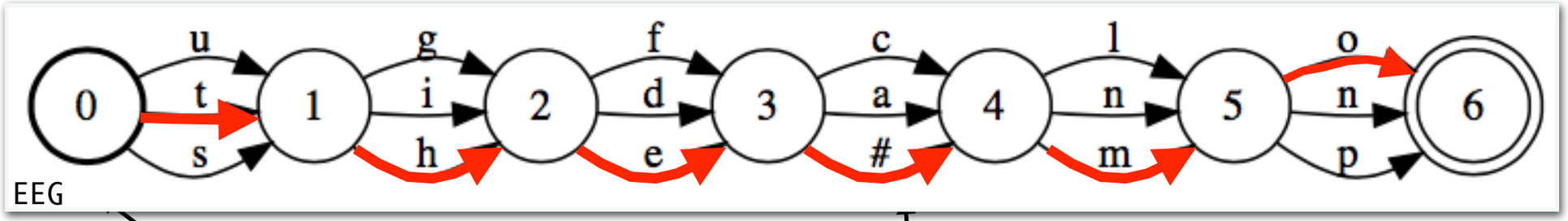

0
1
2
3
4
5
6
u
t
s
g
i
h
f
d
e
c
a
#
l
n
m
o
n
p
EEG

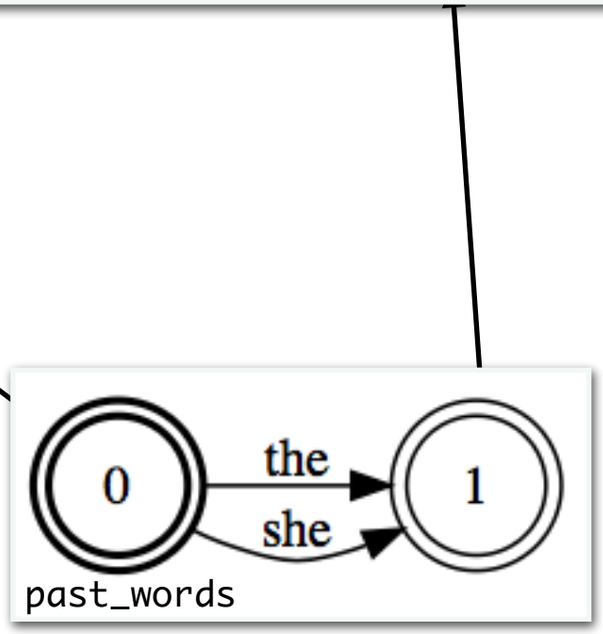

0
s:<epsilon>
5
h:<epsilon>
6
e:<epsilon>
7
#:she
8
t:<epsilon>
1
h:<epsilon>
2
e:<epsilon>
3
#:the
4


u
t
s
g
i
h
f
d
e
c
a
#
l
n
m
o
n
p
0
1
2
3
4
5
6
EEG


s:<epsilon>
t:<epsilon>
h:<epsilon>
h:<epsilon>
e:<epsilon>
e:<epsilon>
#:she
#:the
0
1
2
3
4
5
6
7
8

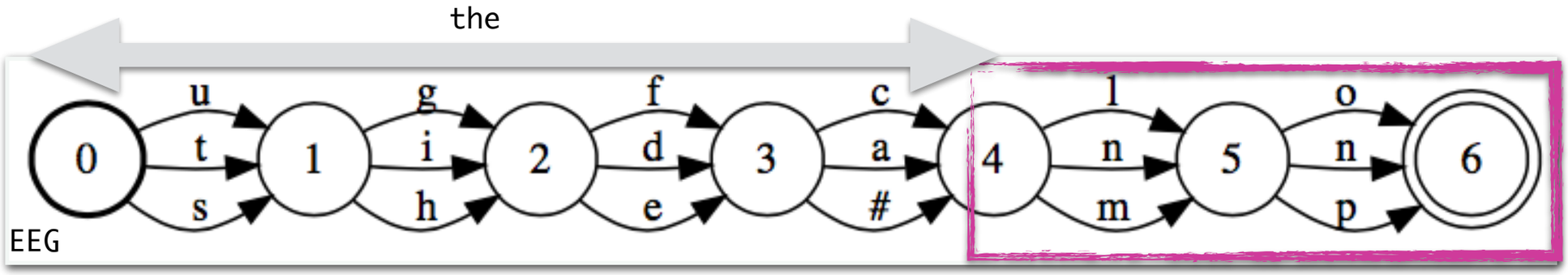
Find possible word history sequence



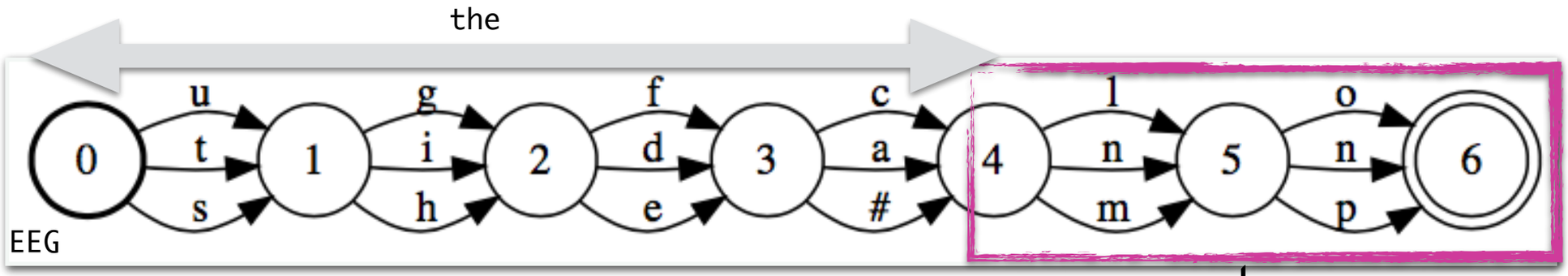


**state 0 is a final state as the entire sequence can be considered a prefix as well

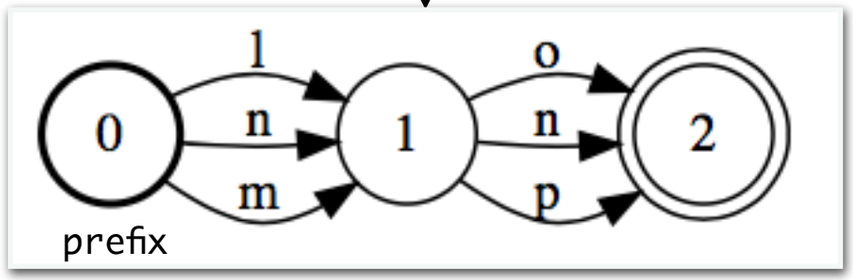
Find possible word completions w/ no context.



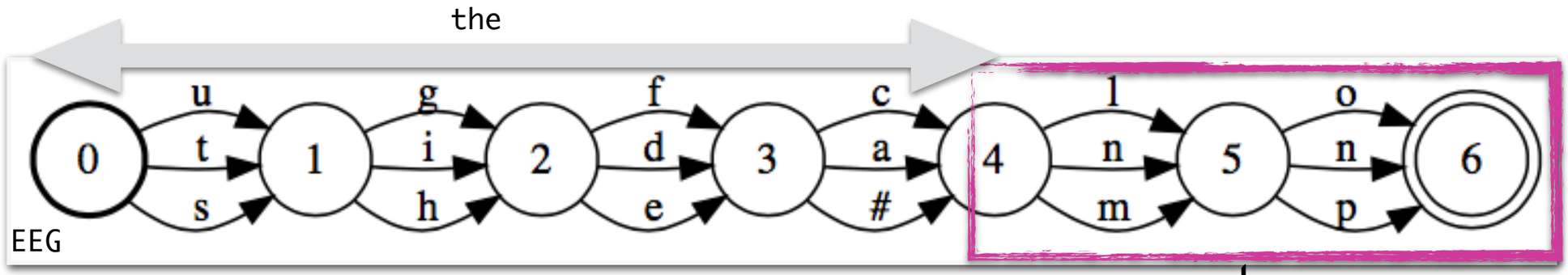
Find possible word completions w/ no context.



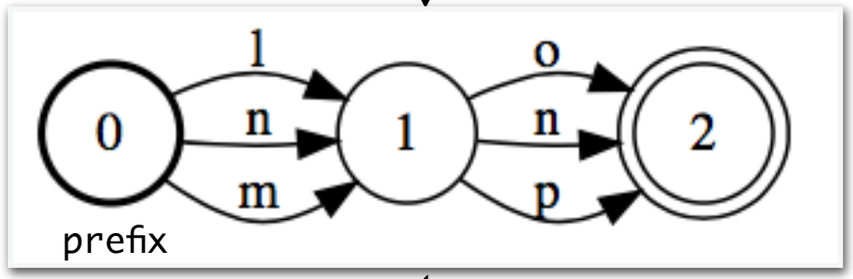
extract the residual prefix



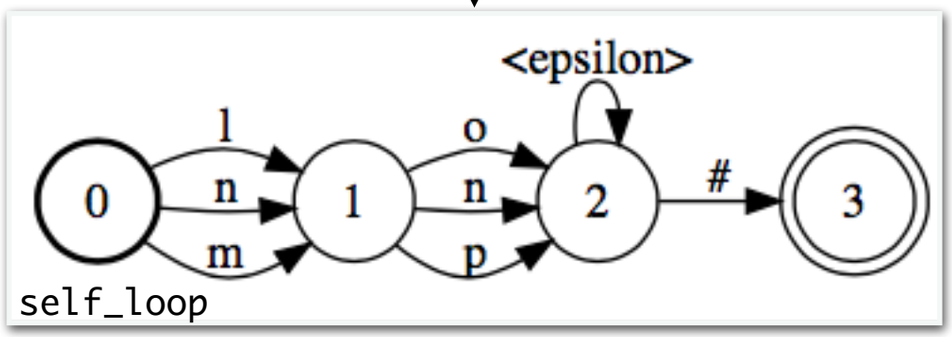
Find possible word completions w/ no context.



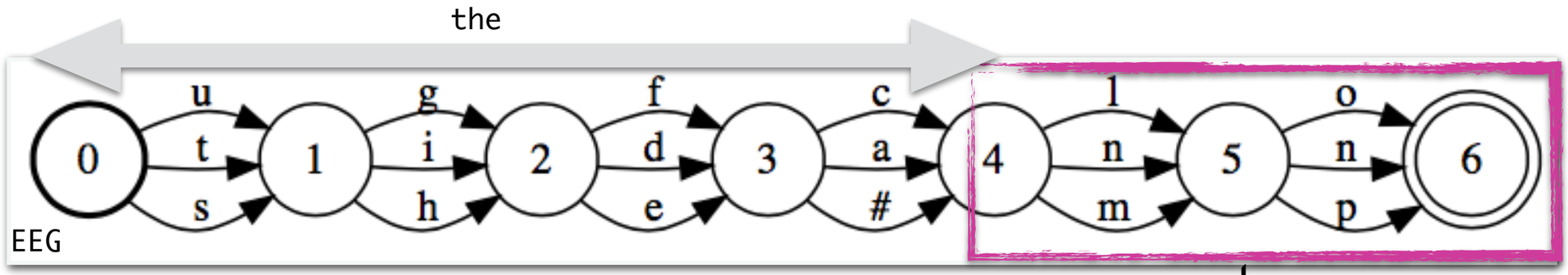
extract the residual prefix



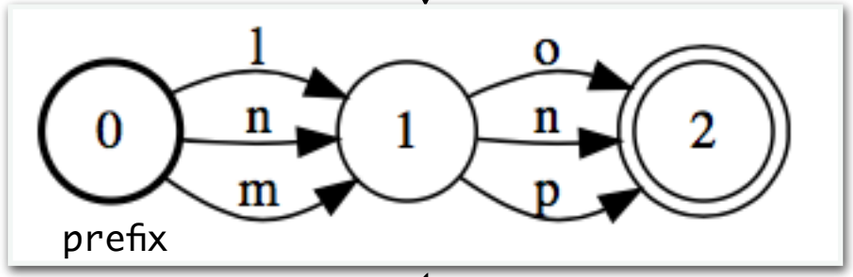
add self-loop



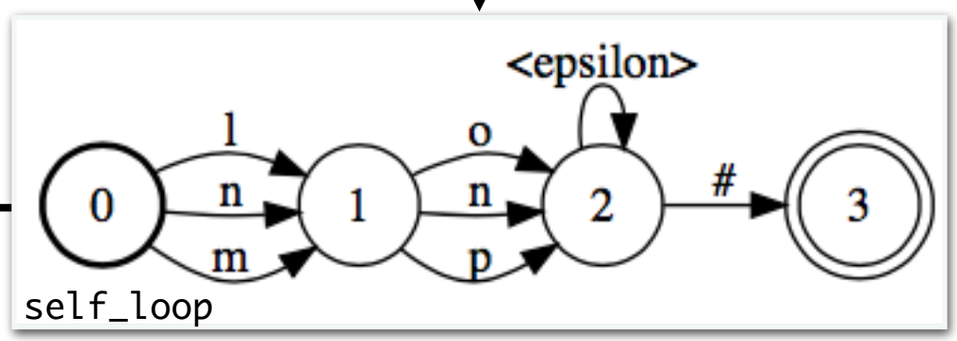
Find possible word completions w/ no context.



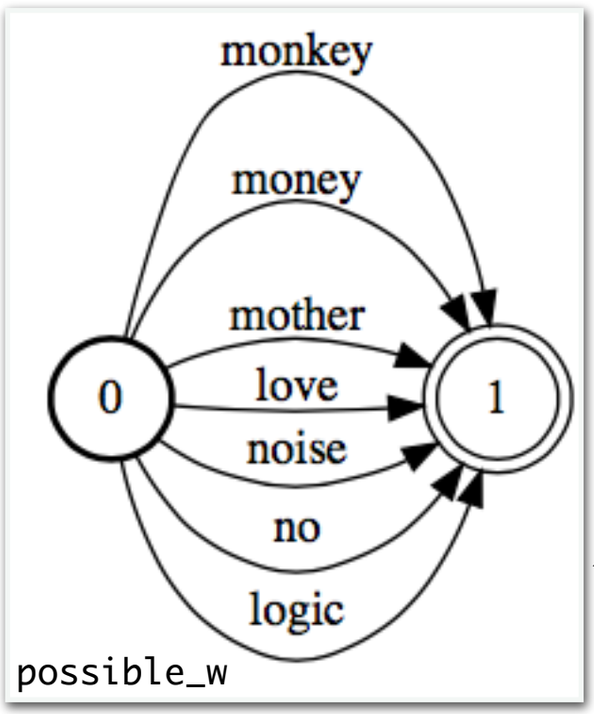
extract the residual prefix



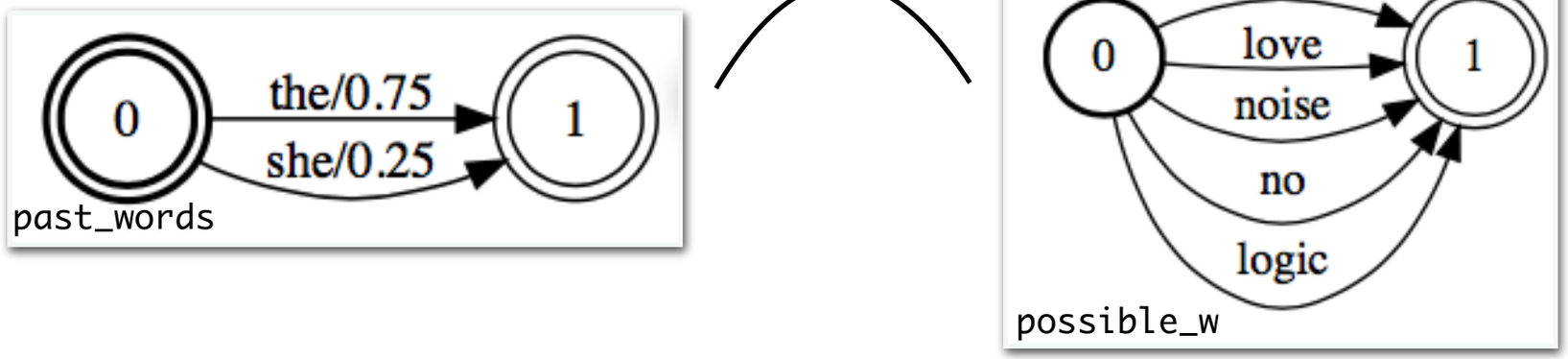
add self-loop



extract possible words (no context)

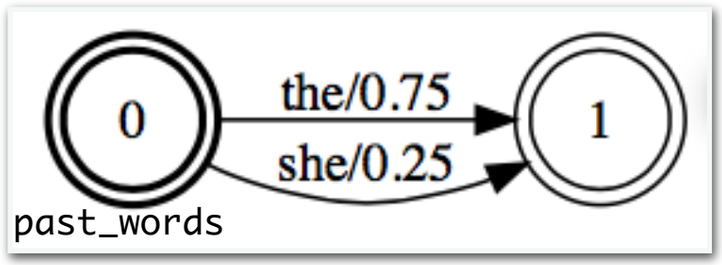


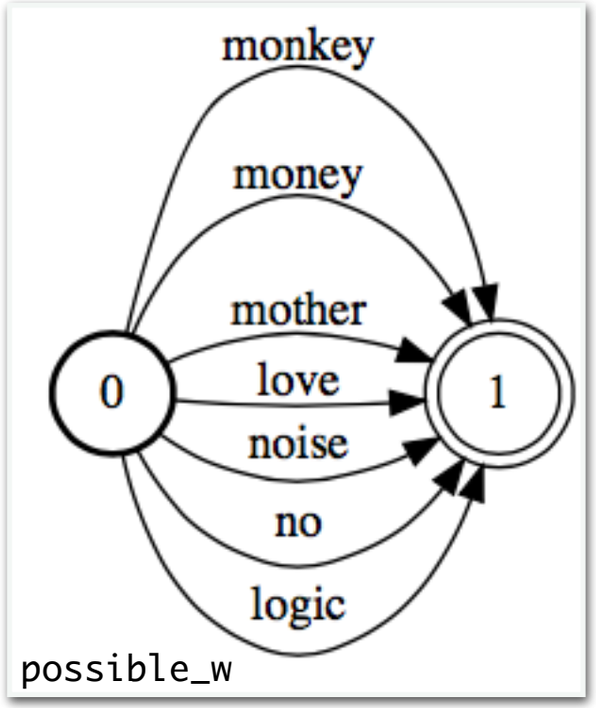
Find top-*k* words the user might be in the middle of typing



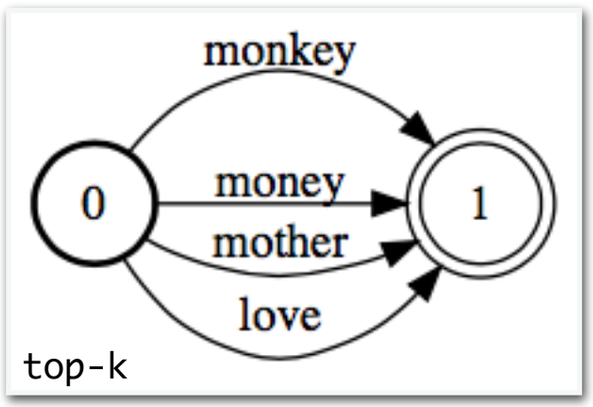
concatenate past and current words

Find top-*k* words the user might be in the middle of typing



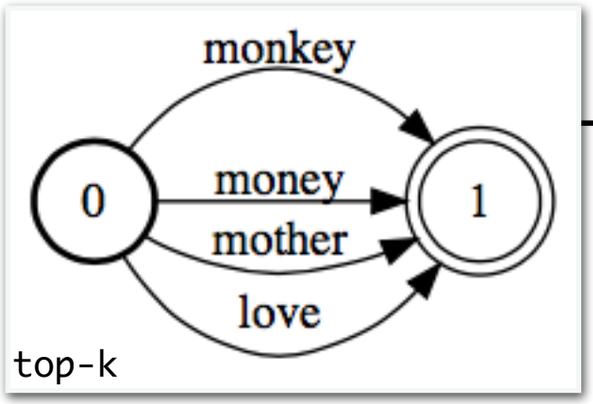


concatenate past and current words

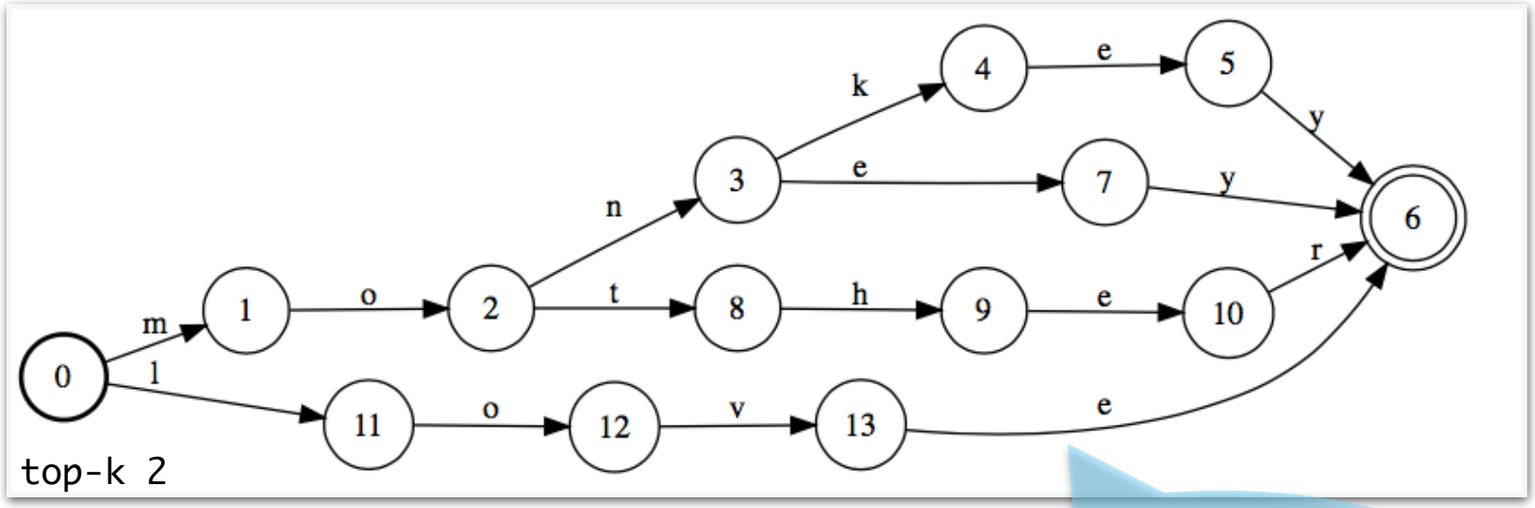


find top-*k* words given context

Make the symbol language model

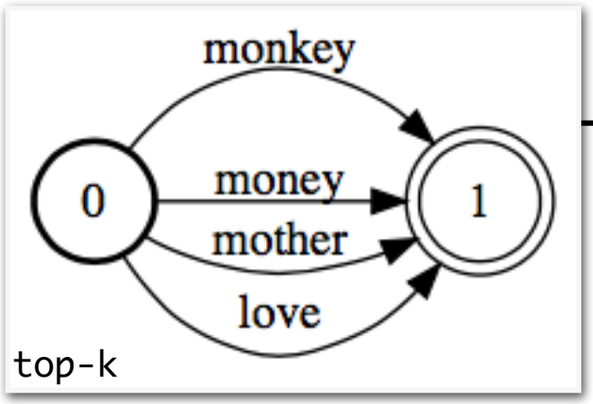


word to symbol

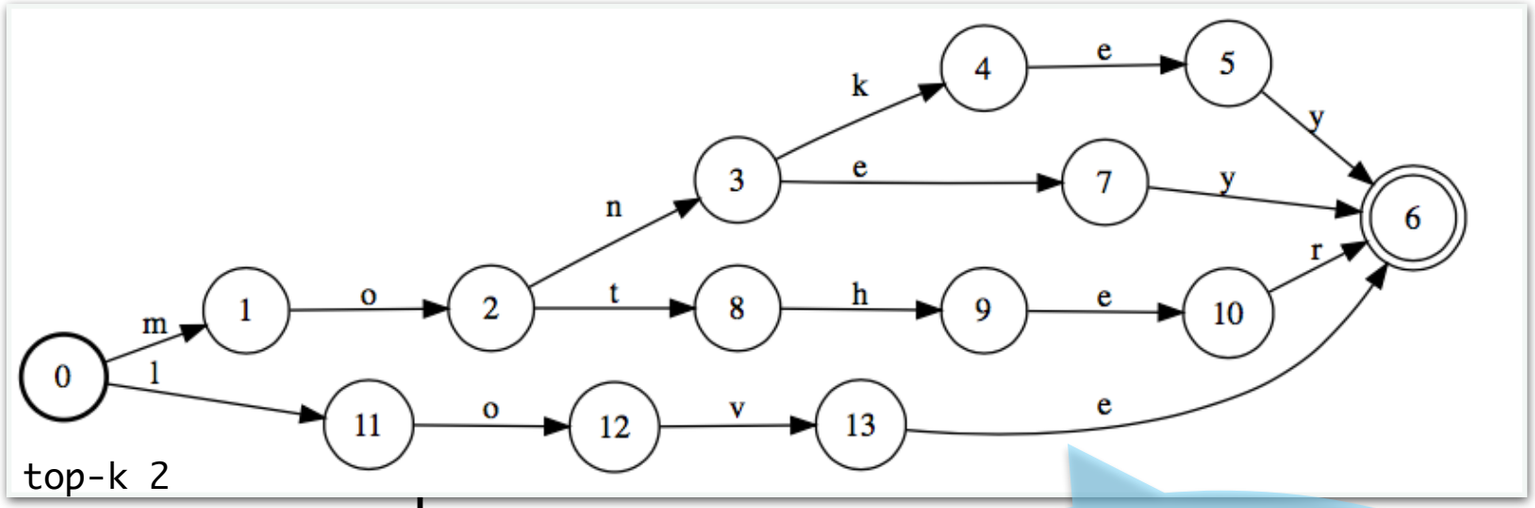


symbol probabilities reflect word probabilities.

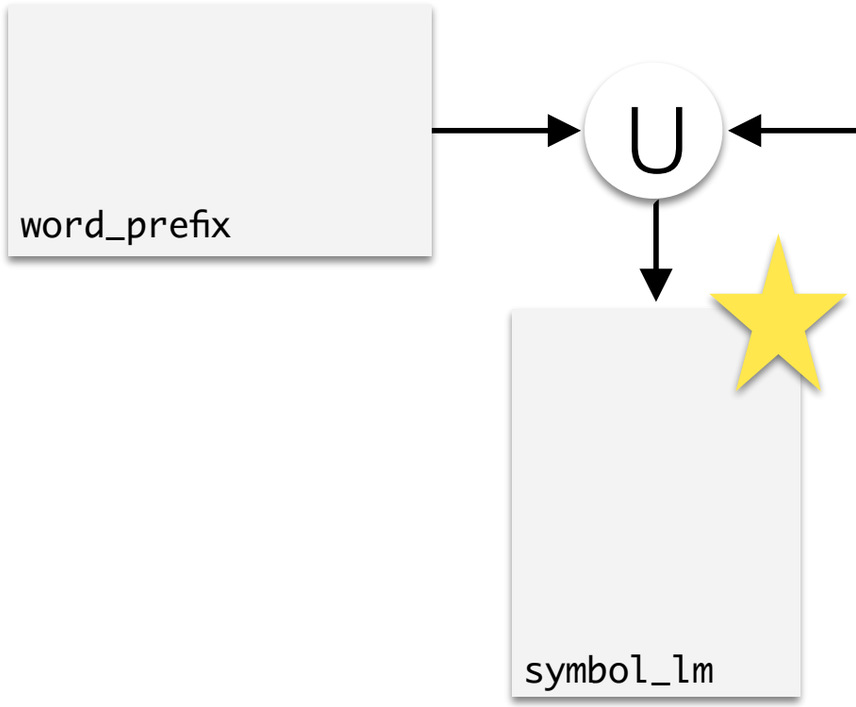
Make the symbol language model



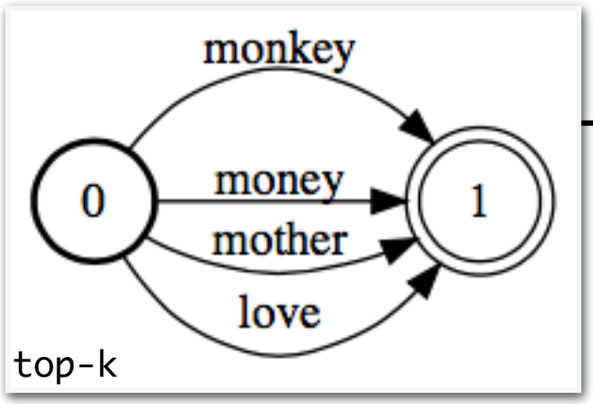
word to symbol



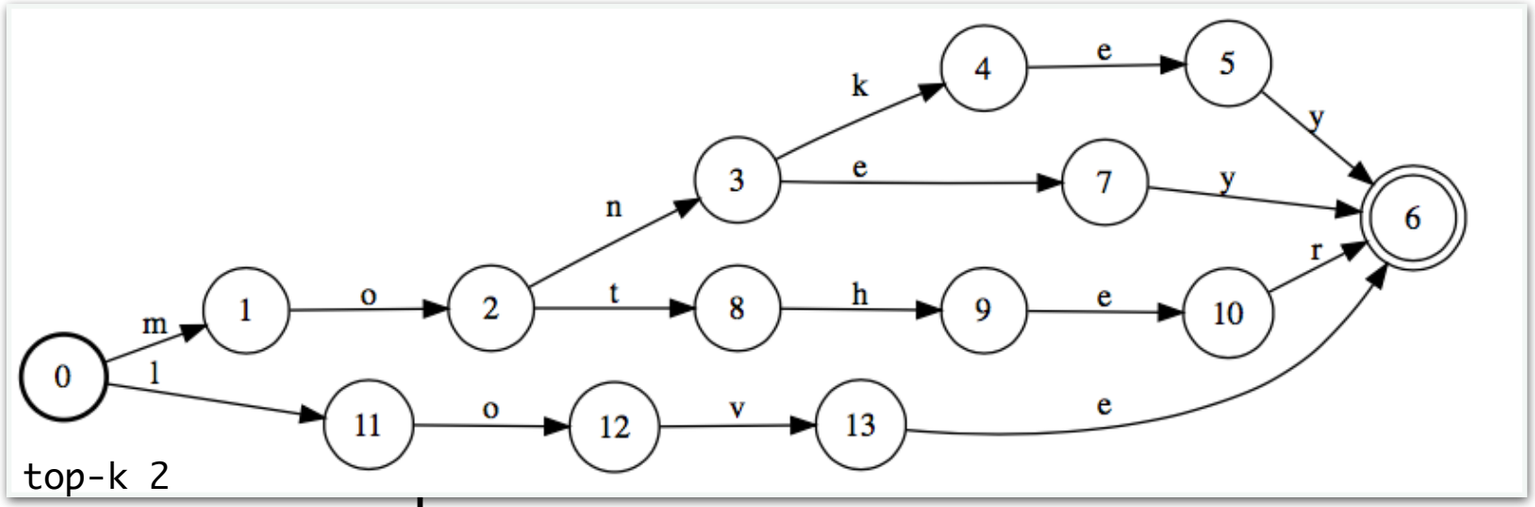
symbol probabilities reflect word probabilities.

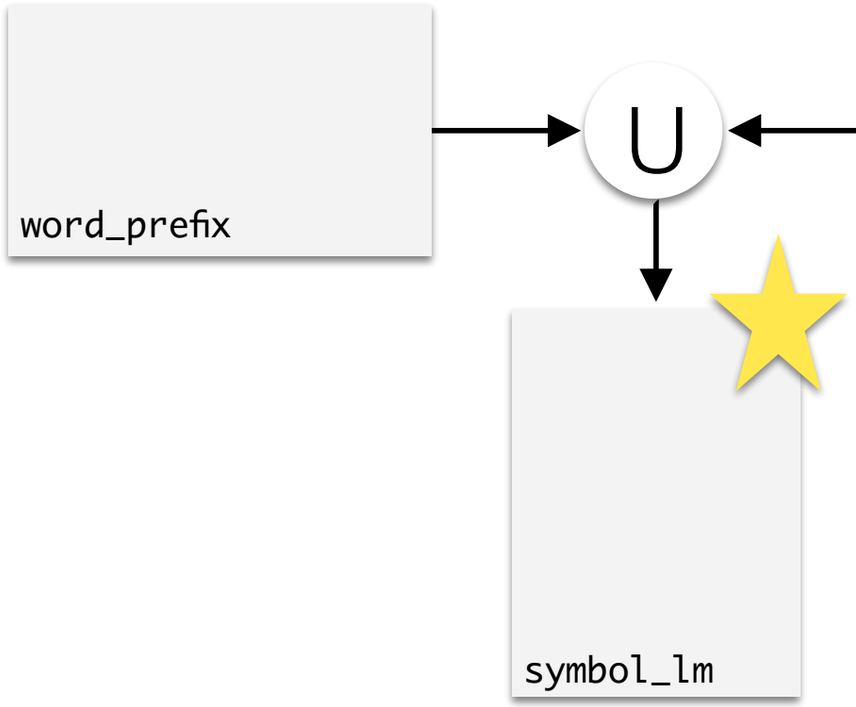


Make the symbol language model

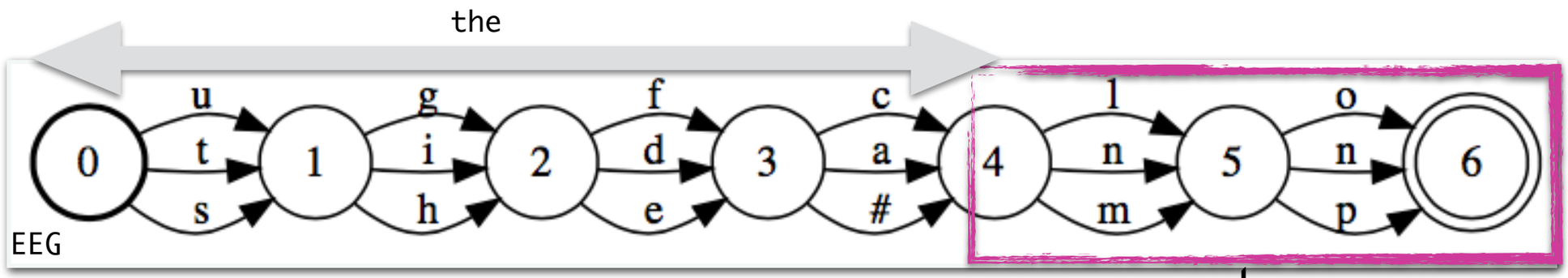


word to symbol

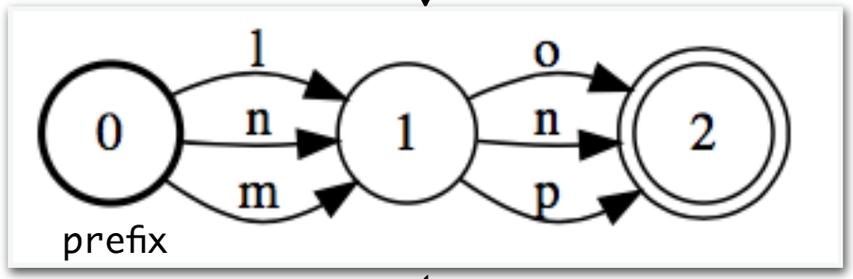


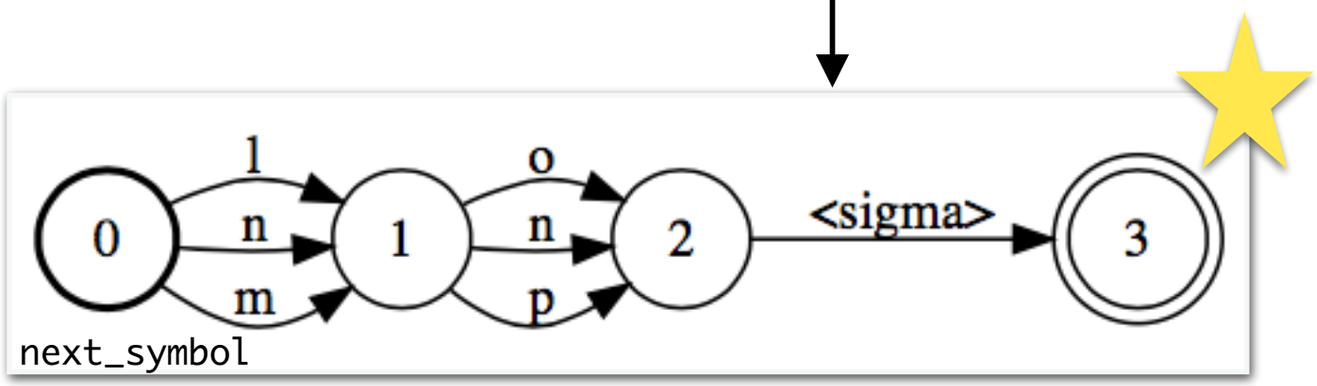


Prepare the current trailing prefix to find its next symbol distribution

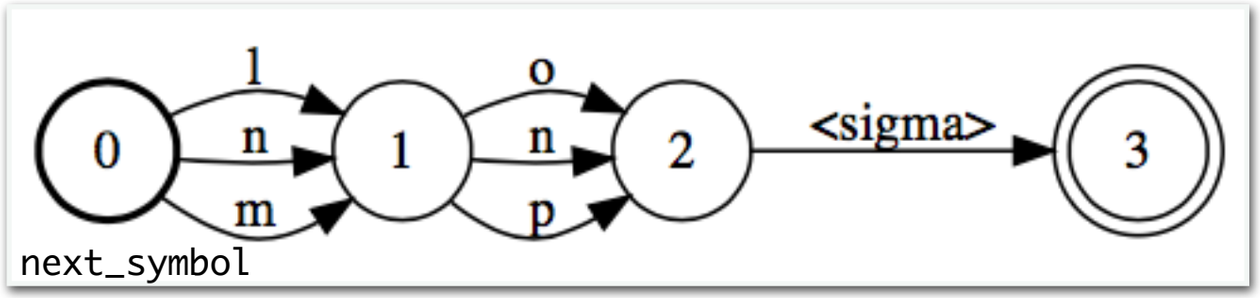


extract the residual prefix





Intersect symbol_lm (the machine in #3) with next_symbol (the machine in #4) and get the distribution over final symbols to return to front-end.



X

symbol_lm

symbol	probability
n	0.103
t	0.102
g	0.09
o	0.09
a	0.06
r	0.04
s	0.04
v	0.01

OCLM in short

1. Extract the potential previous words and possible current ones to figure out the current word and create a targeted LM
2. Extract the potential prefix
3. Extract the next symbol in the potential prefix given the LM

symbol/letter prediction with word knowledge to improve letter prediction

$$P(sym|EEG) \propto P(sym)P(EEG|sym)$$

$$P(sym) = P(sym|X_m)$$

- ✓ $X_0 = \{sym_{t-i}\}, i > 0$
- ✓ $X_1 = \{word_{t-i}\}, i > 0$
- ✓ $X_2 = \{EEG_{t-i}\}, i \geq 0$

Candidate symbols

symbol: confidence

u: 0.95
m: 0.05

Evaluation of OCLM

Method:

Train different LM types:

1) 5 gram LM applied on basic LM algorithm

2) Prefix LM applied on basic algorithm*

3) OCLM algorithm

80% train 20% test, on Brown corpus

Evaluation Metric #1

Mean Reciprocal Rank (MRR):

Or how close were we in our guess for target symbol?

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Evaluation Metric #2

Perplexity:

Or how likely are we to reproduce the data with our model

$$PPL(t, x) = 2^{-\frac{1}{T}\sum_{t=1}^{T} \log q_{model}(x{=}target_t)}$$

Evaluation Metric #3

ACC@10:

Or how often the target was in top 10 guesses

metric	NGRAM	PreLM	OCLM
MRR	0.4	0.7	0.75
PPX	4.4	1.8	1.9
ACC@10	0.69	0.96	0.96

Table 1: Evaluation Results (n=1)

nbest	metric	PreLM	OCLM
	MRR	0.29	0.51
n=2	PPX	3.5	3.0
	ACC@10	0.69	0.87
	MRR	0.26	0.44
n=3	PPX	4	3.9
	ACC@10	0.63	0.83

Table 2: Evaluation Results (n=2, n=3)

Conclusions

Both OCLM and PreLM outperform NgramLM in terms of PPX and MRR on a single EEG evidence

On multiple hypotheses both algorithms degraded but OCLM was performing much better than PreLM

OCLM's shortcoming: long runtime

Additional material

The research paper: A Multi-Context Character Prediction Model for a Brain-Computer Interface, SCLeM, ACL, 2018

The followup research paper: Noisy Neural Language Modeling for Typing Prediction in BCI Communication, SPLAT Workshop, NAACL, 2019

The git repo: https://github.com/shiranD/oclm

Acknowledgements:

OHSU team:

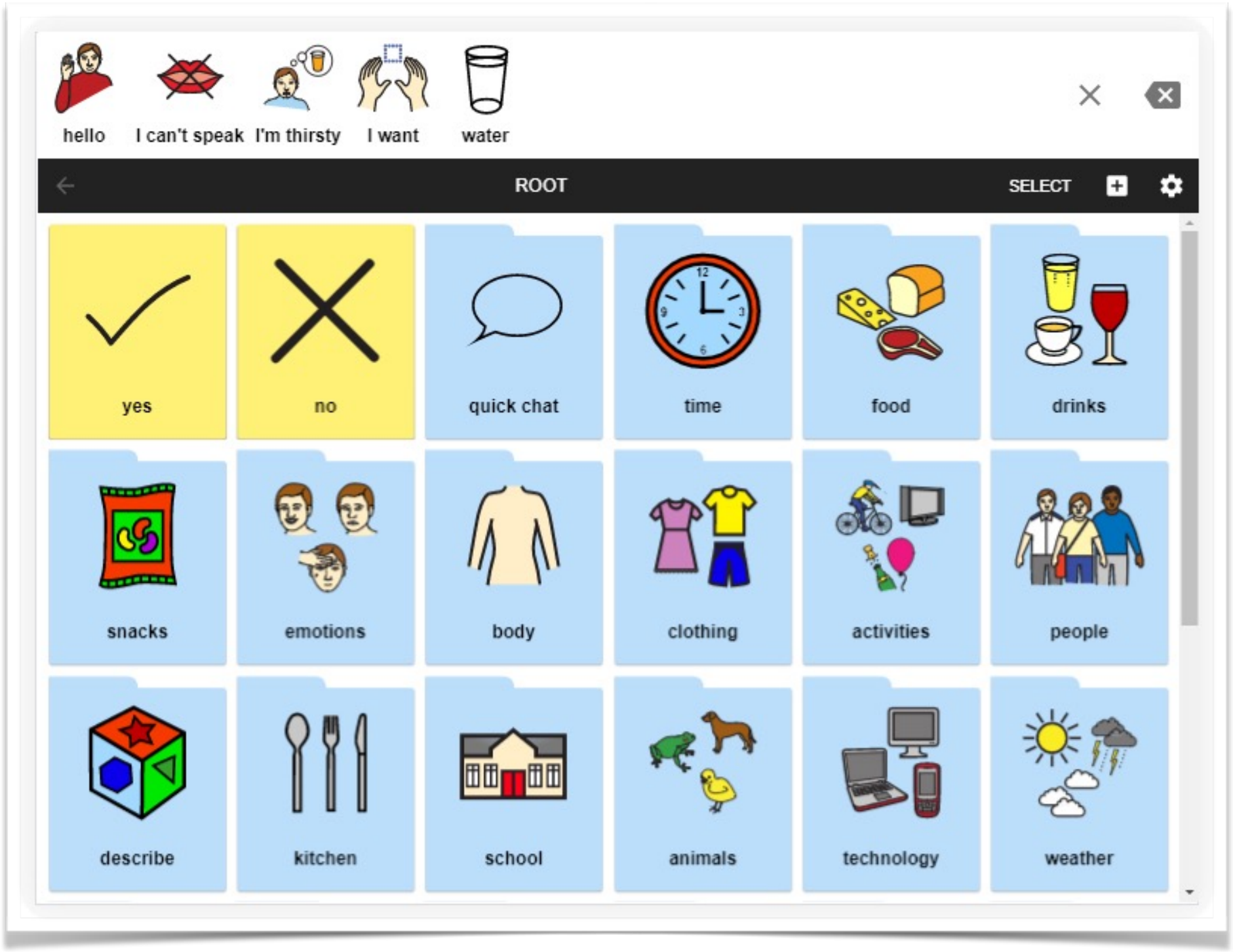
Melanie Fried-Oken
Betts Peters
Barry Oken

Northeastern University team:

David Smith
Shaobin Xu
Rui Dong

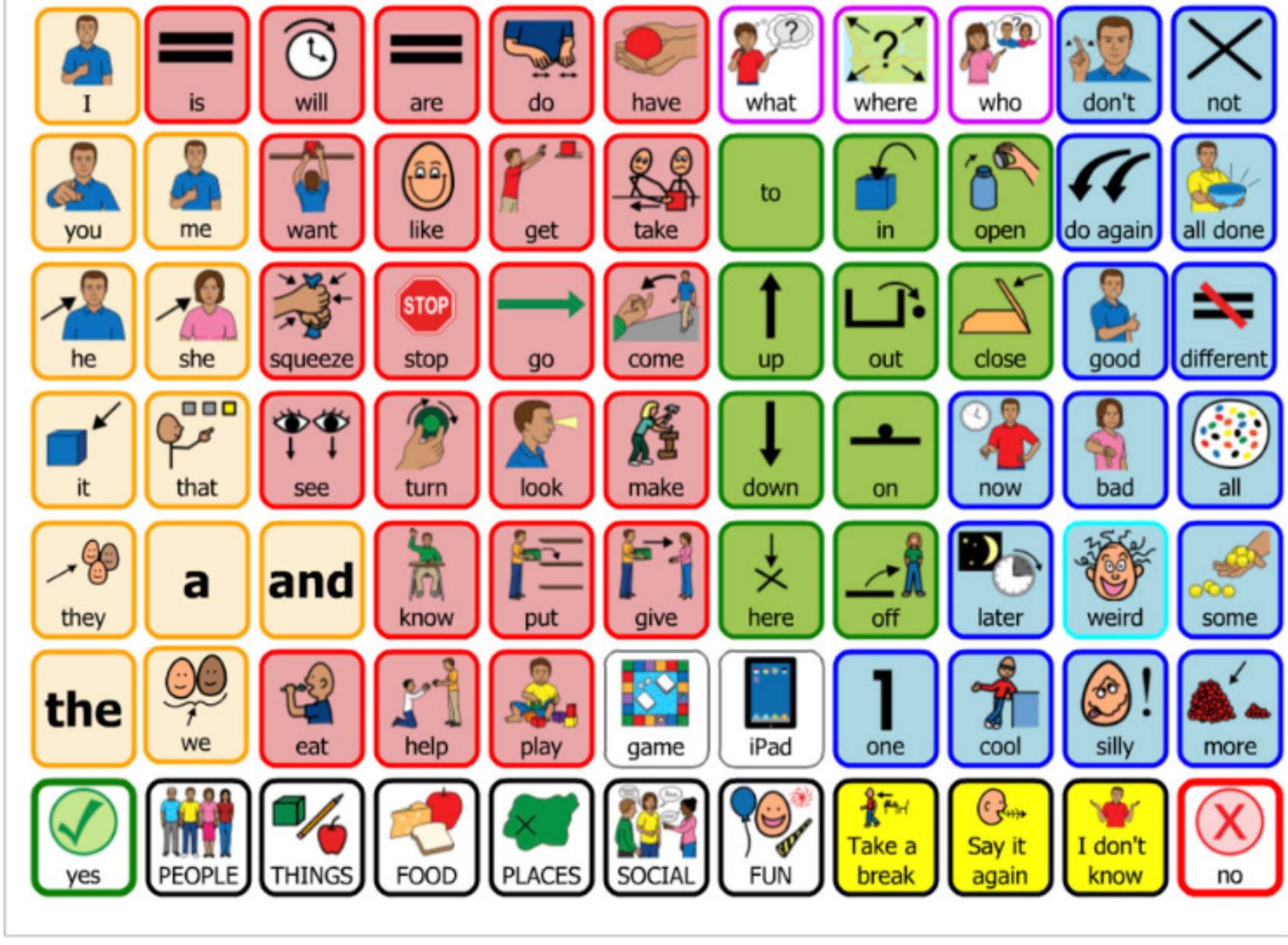
My advisor: Steven Bedrick

Compositional Language Modeling for Icon-Based Augmentative and Alternative Communication



Open-source communication board
https://www.cboard.io/cboard/open-source/2017/12/05/open-source-communication-board/

Core Word Communication Board

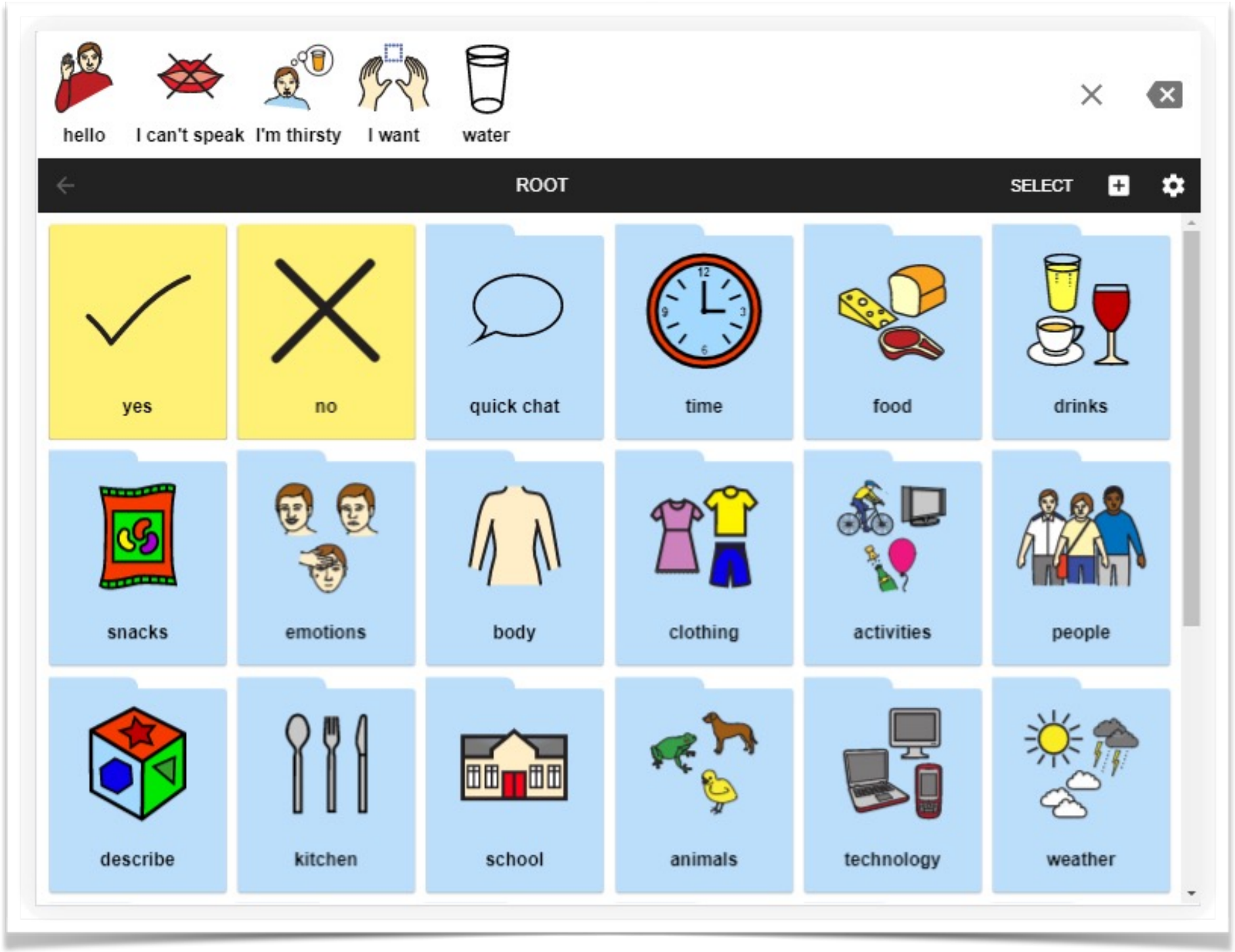


electronic communication board
https://www.rachelmadel.com

Scaling?

Selection modality

Endless tree navigation…



Open-source communication board
https://www.cboard.io/cboard/open-source/2017/12/05/open-source-communication-board/

Limited message complexity…

electronic communication board
https://www.rachelmadel.com

Our goal:

Creating an icon language-model based AAC system

Symbolstix Icon set

Used by communities who are in need of icon-based communication

Symbolstix Icon set

Used by communities who are in need of icon-based communication

Human Curated icons

Symbolistix Icon set

Used by communities who are in need of icon-based communication

Human Curated icons

34,837 icons: 13,951 single words, 12,434 unique single words

name: agree

word type: verb

synonyms: agreement

agreed

agrees

agreeing

approve

flexibility

concur

on the same page

see eye to eye

name: agree

word type: verb

synonyms: agreement

agreed agrees

agreeing approve

flexibility concur

on the same page

see eye to eye

name: agree

word type: verb

synonyms: agreement

agreed agrees

agreeing approve

flexibility concur

on the same page

see eye to eye

Symbolistix Icon set

Used by communities who are in need of icon-based communication

Human Curated icons

34,837 icons: 13,951 single words, 12,434 unique single words

Symbolistix Icon set

Used by communities who are in need of icon-based communication

Human Curated icons

34,837 icons: 13,951 single words, 12,434 unique single words

No Corpus available!

Our Question:

How to create language models for corpus-less symbol-set

We don’t have:

We don’t have:

icon corpus

We don’t have:

~~icon corpus~~

We don’t have:

~~icon corpus~~

we have:

We don’t have:

~~icon corpus~~

we have:

- Icon meta-data

We don’t have:

~~icon corpus~~

we have:

- Icon meta-data
- pre existing **textual** corpora such as Gigaword

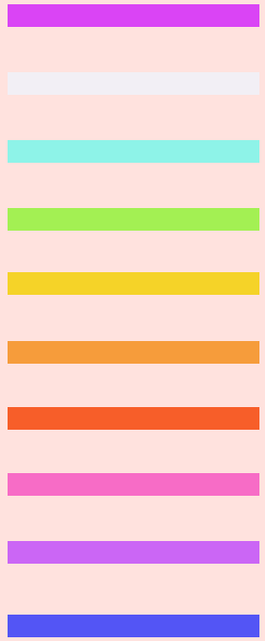
We don’t have:

~~icon corpus~~

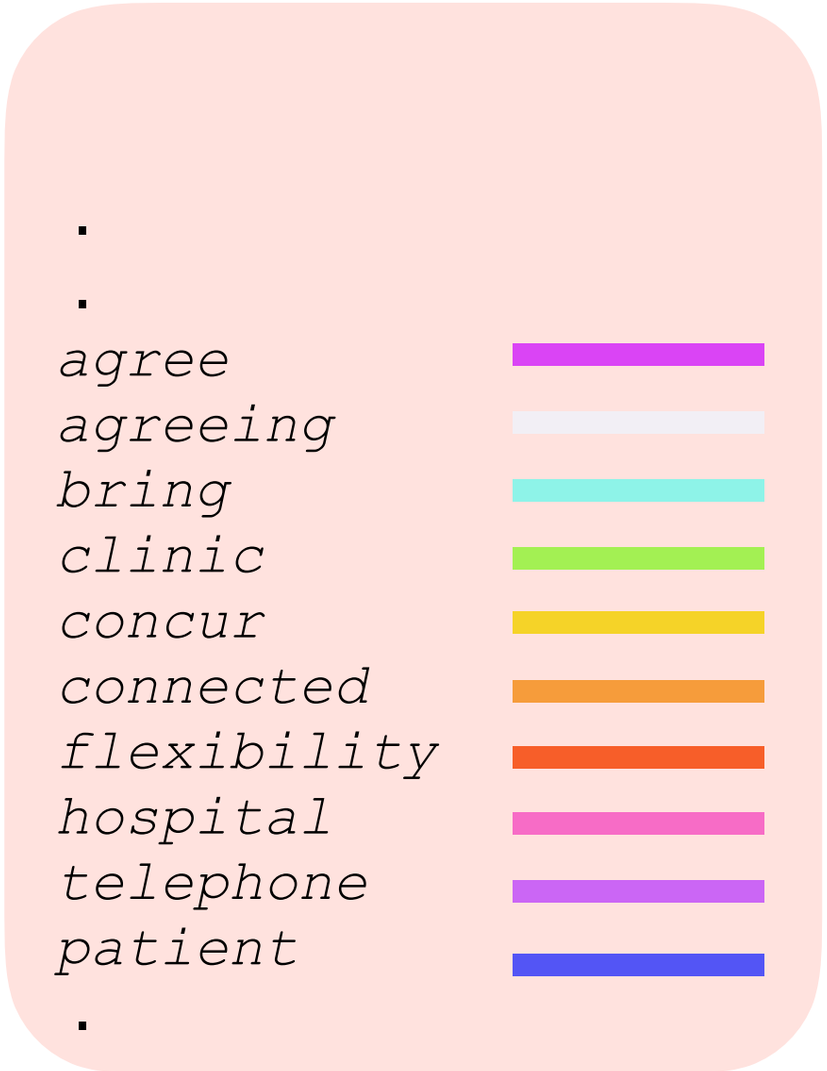
we have:

- Icon meta-data
- pre existing textual corpora such as Gigaword
- Word embedding representation

.
.
agree
agreeing
bring
clinic
concur
connected
flexibility
hospital
telephone
patient
.



telephone



concur

agree

agreeing

patient

clinic

hospital

telephone

concur

agree

agreeing

.
.
agree
agreeing
bring
clinic
concur
connected
flexibility
hospital
telephone
patient
.

patient

clinic

hospital

We don’t have:

~~icon corpus~~

we have:

- Icon meta-data
- pre existing textual corpora such as Gigaword
- Word embedding representation

Perhaps we can utilize the synonyms, apply to textual corpus?

Icon representation

name: agree

word type: verb

synonyms: agreement

agreed

agrees

agreeing

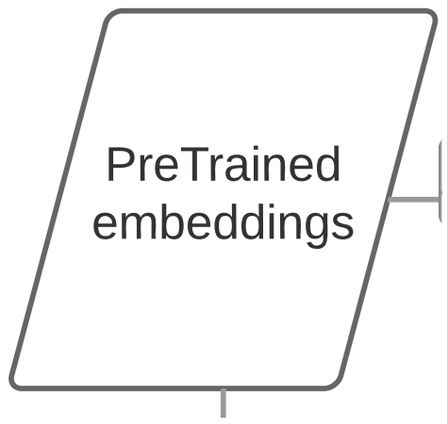
approve

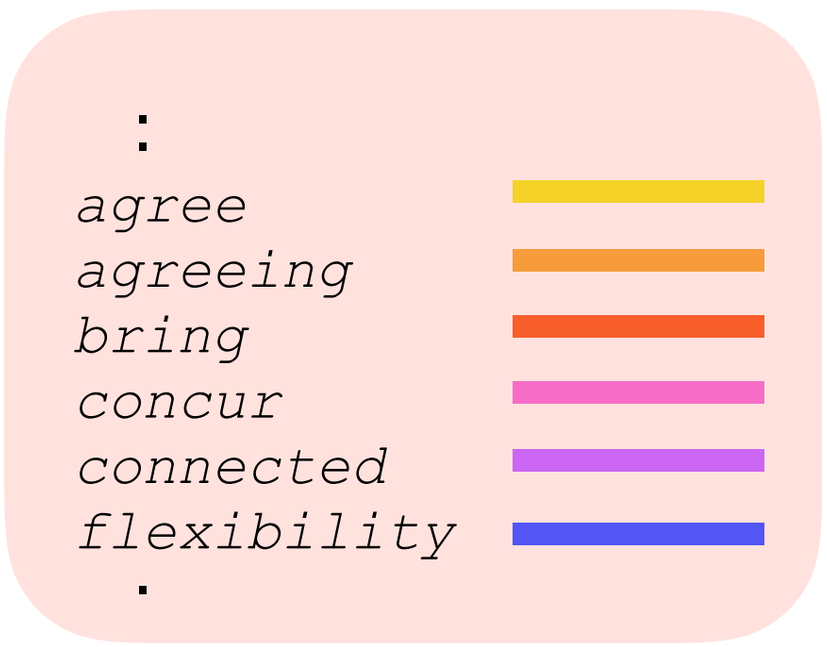
flexibility

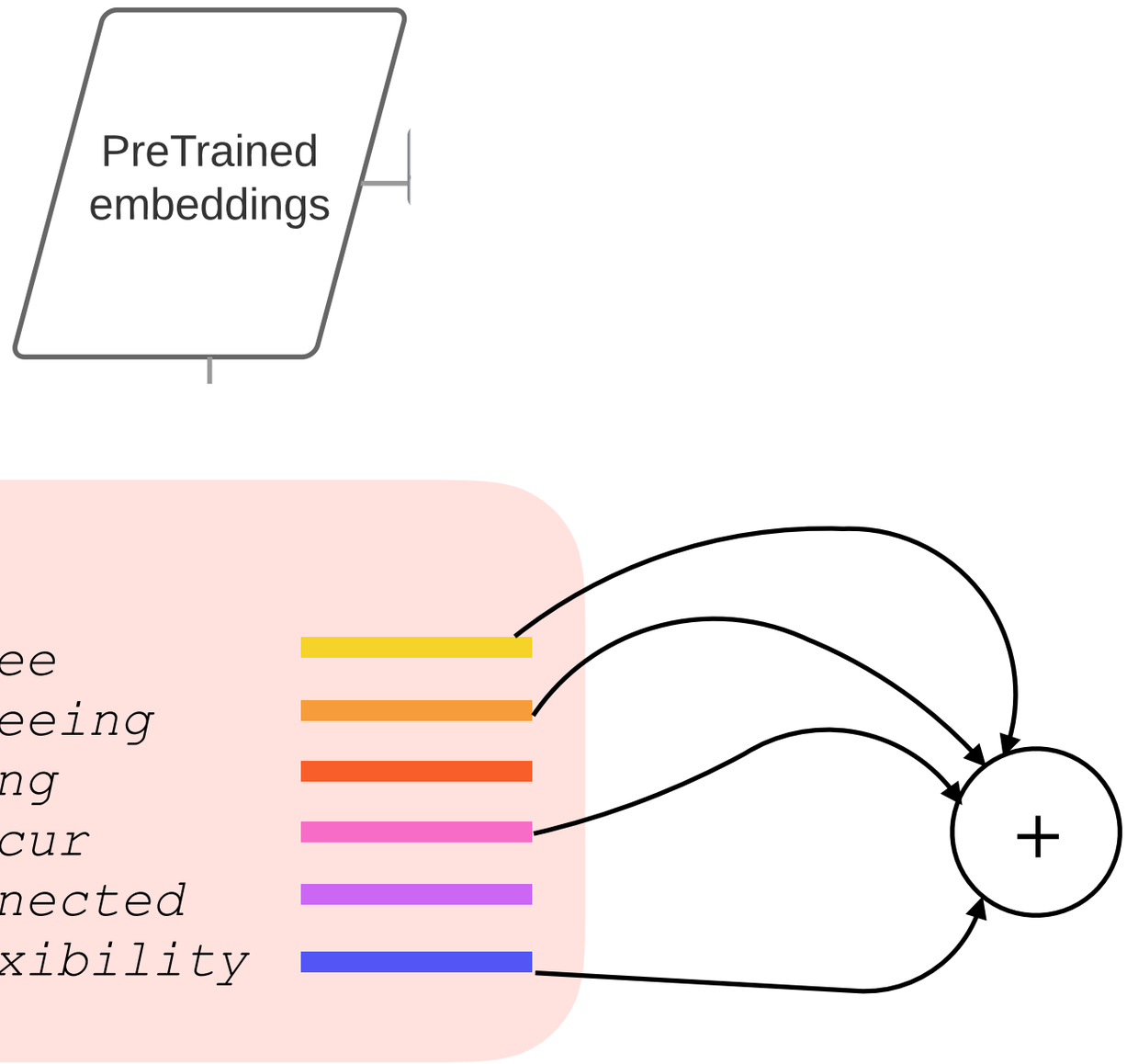
concur

on the same page

see eye to eye


PreTrained
embeddings


:
agree
agreeing
bring
concur
connected
flexibility
.

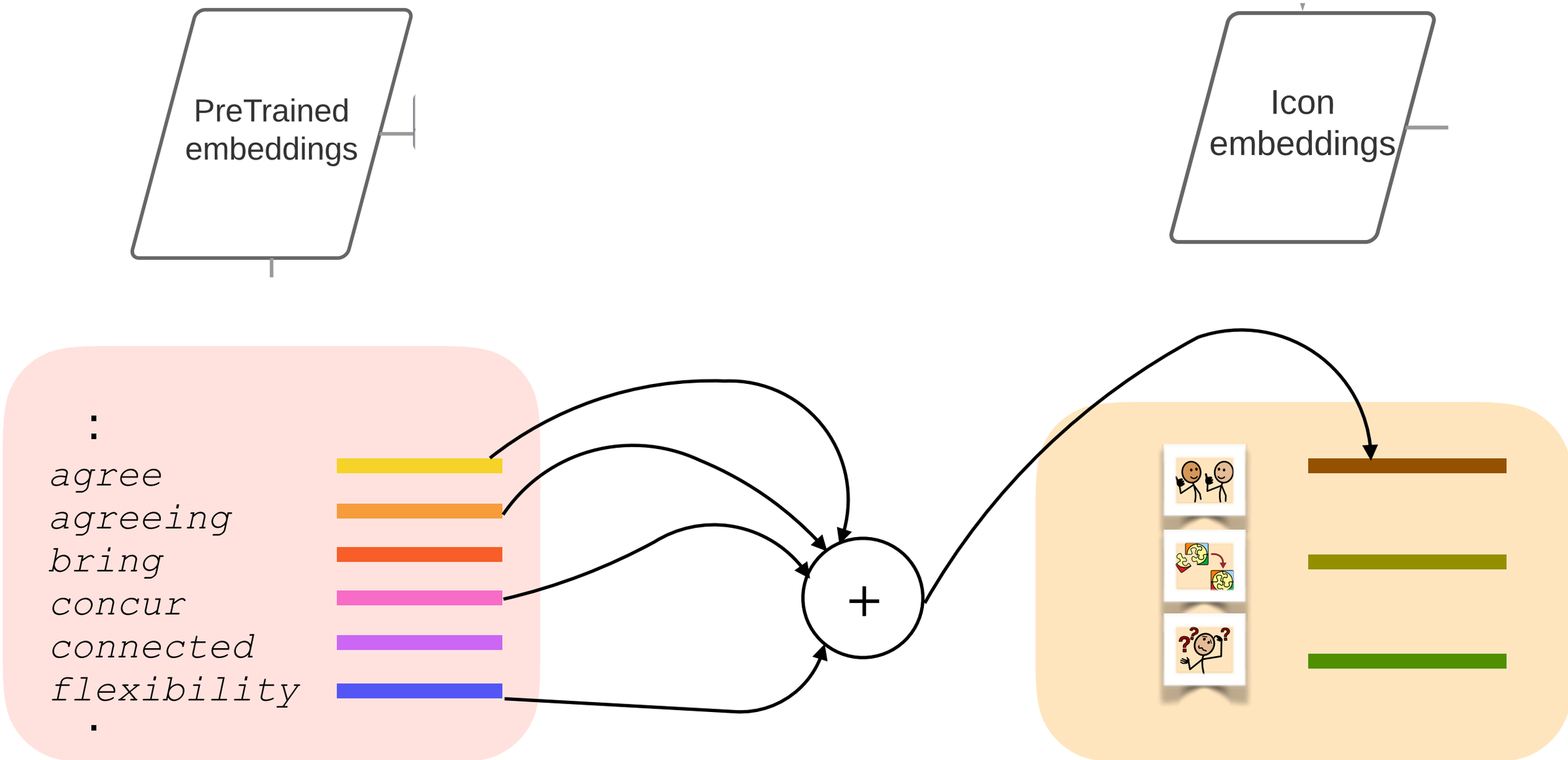

PreTrained
embeddings
:
agree
agreeing
bring
concur
connected
flexibility
.
+

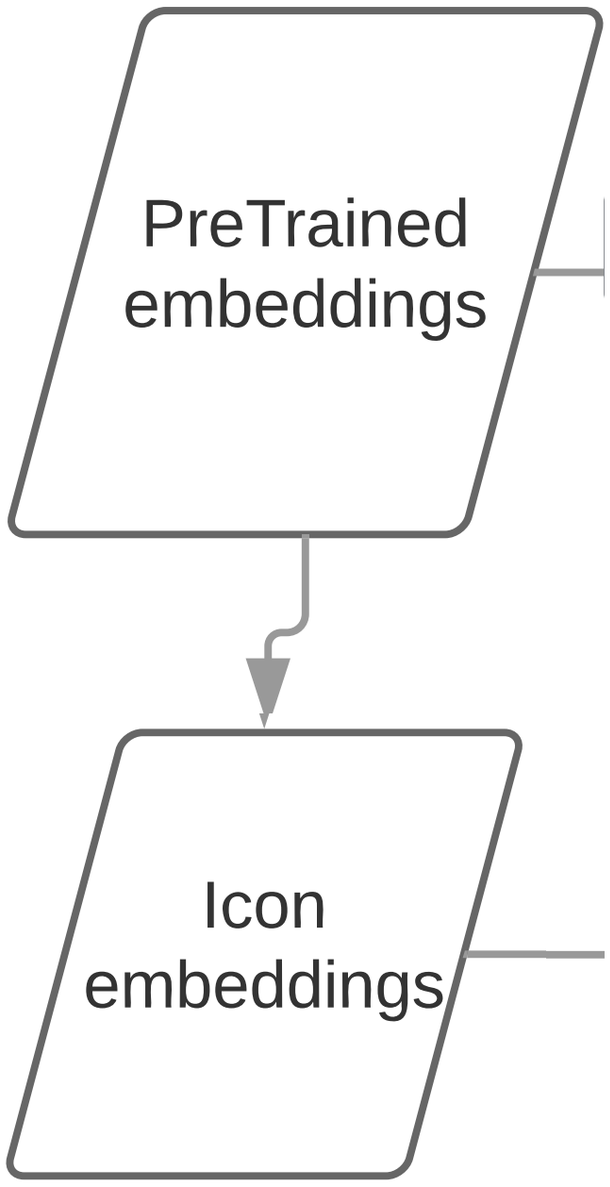
PreTrained embeddings

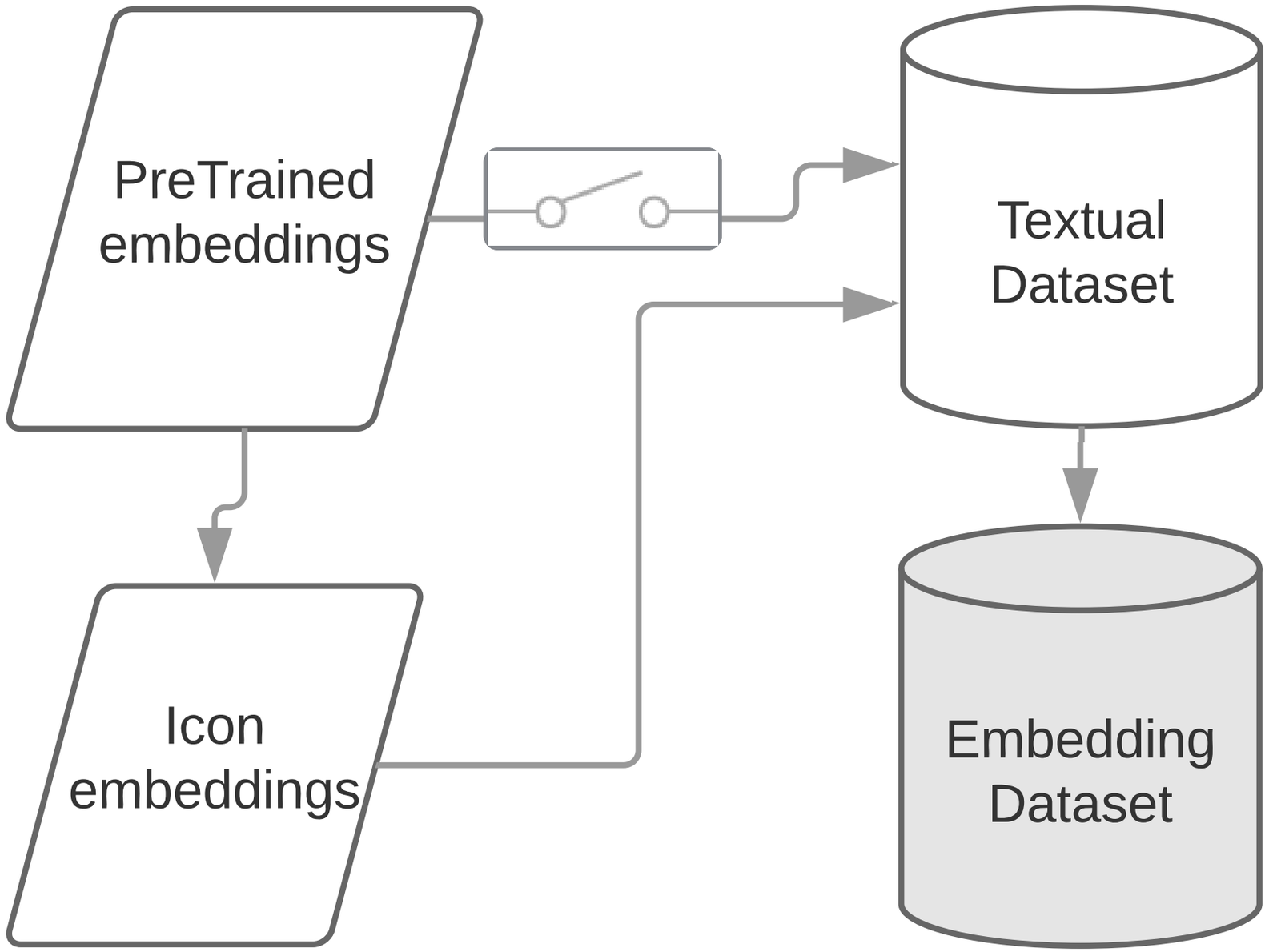
Icon embeddings

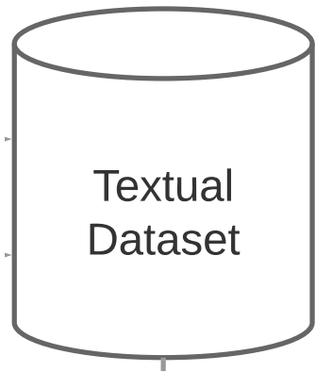
:
agree
agreeing
bring
concur
connected
flexibility
.

+

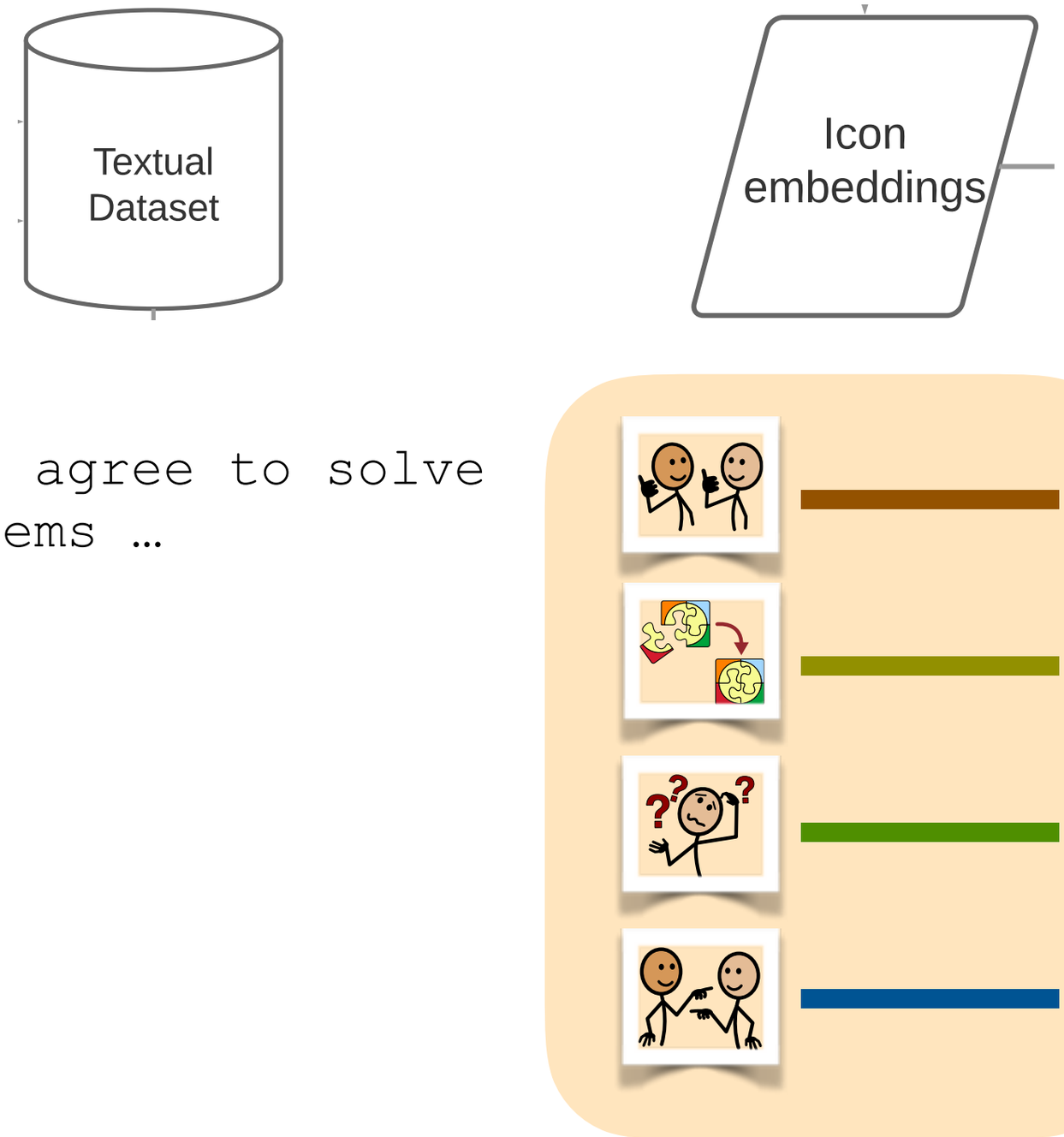

PreTrained embeddings
Icon embeddings
:
agree
agreeing
bring
concur
connected
flexibility
.
+


PreTrained
embeddings
Icon
embeddings

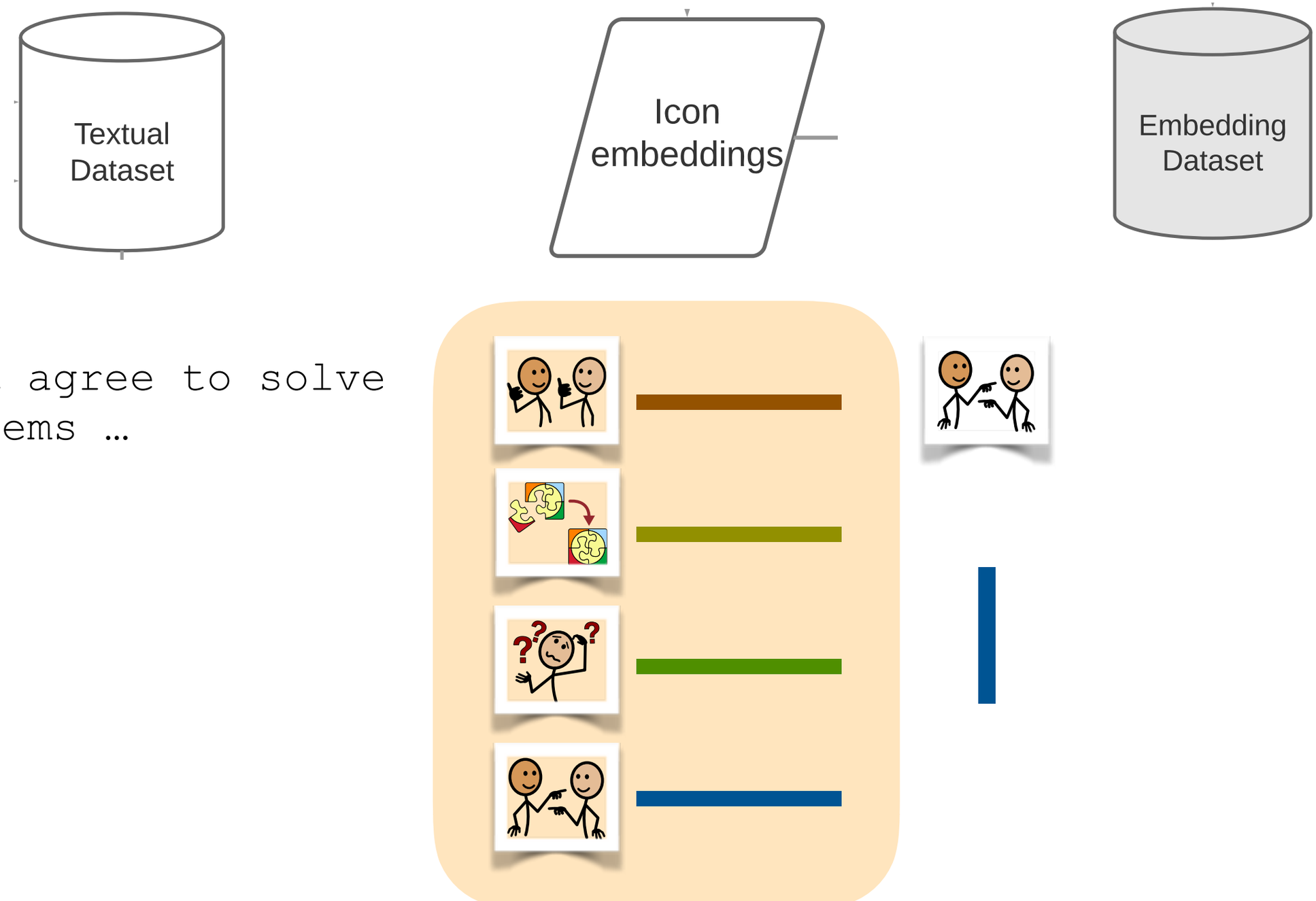

PreTrained
embeddings
Textual
Dataset
Icon
embeddings
Embedding
Dataset



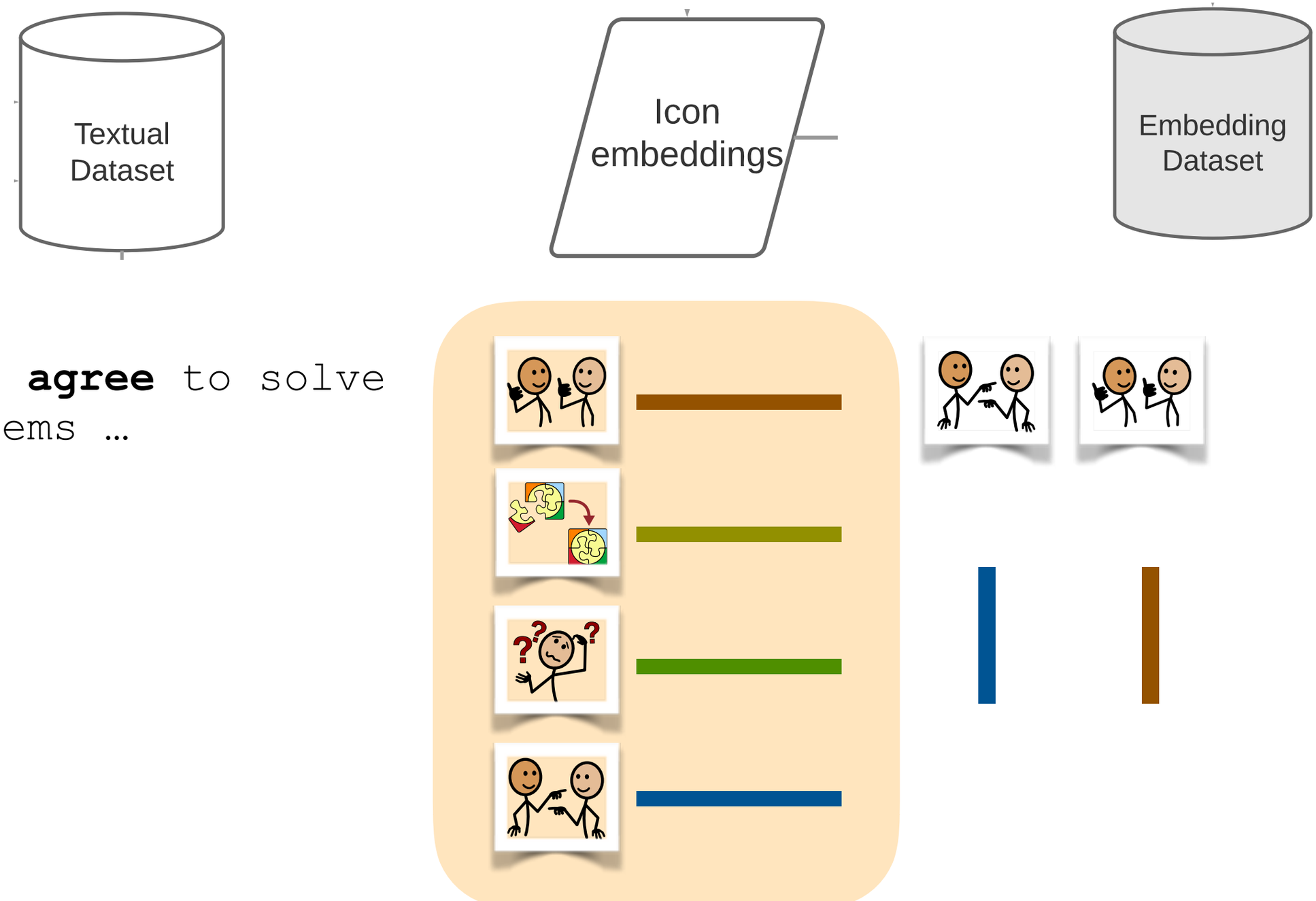
… you agree to solve problems …



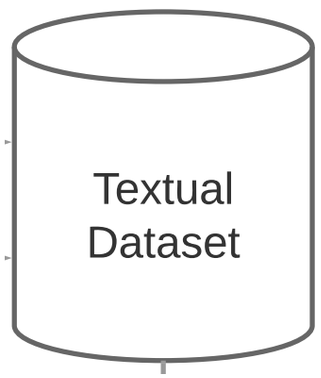
… you agree to solve problems …

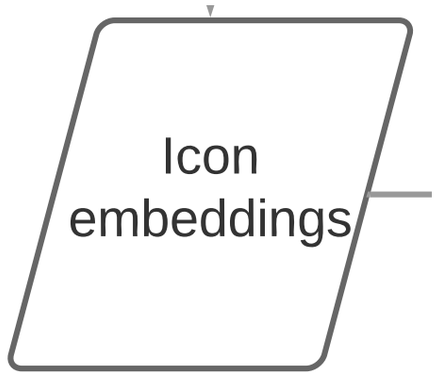


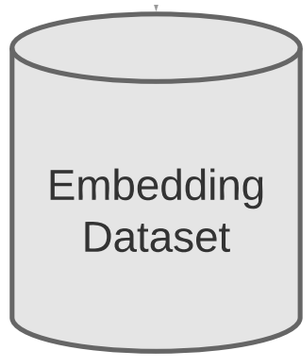
… **you** agree to solve problems …



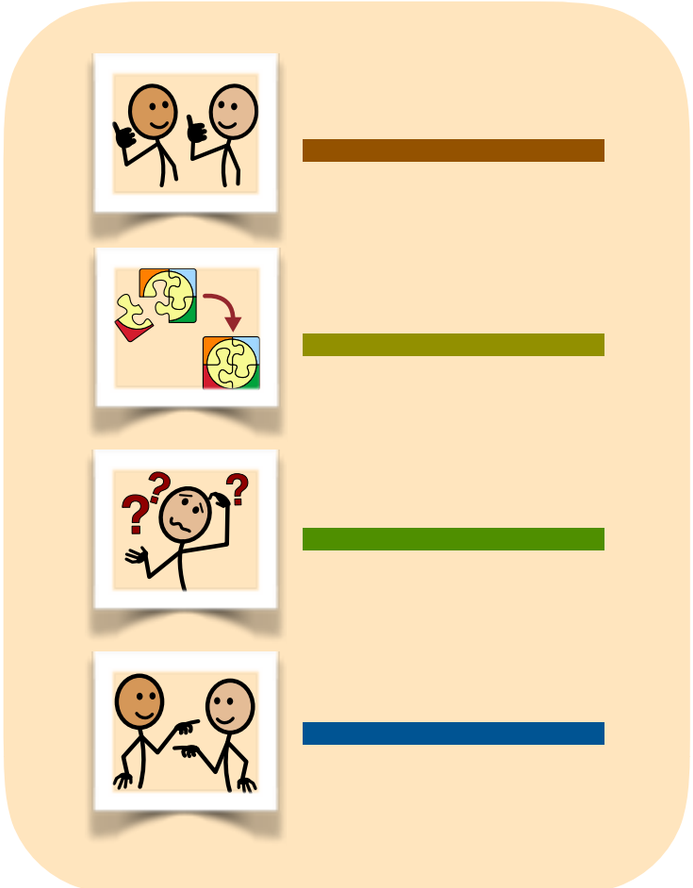
… you **agree** to solve problems …

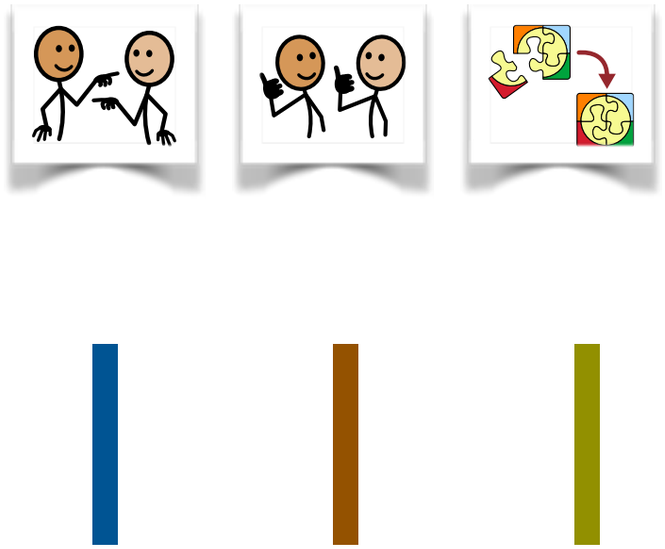


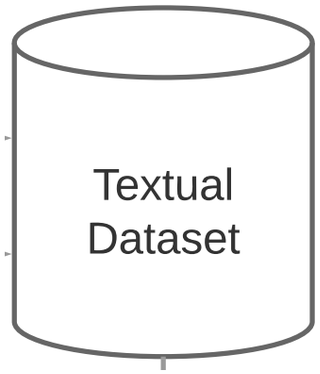


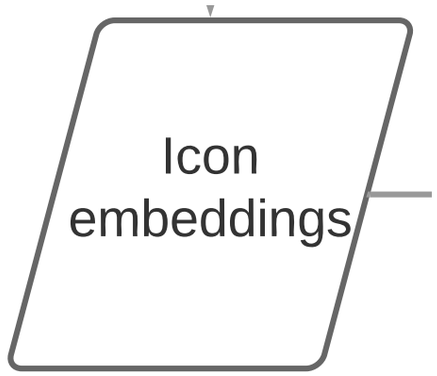


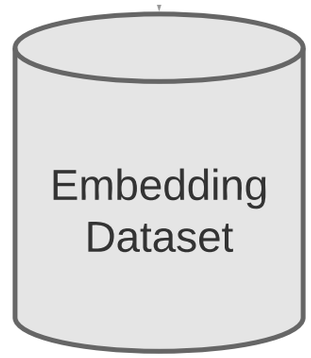
… you agree to **solve** problems …



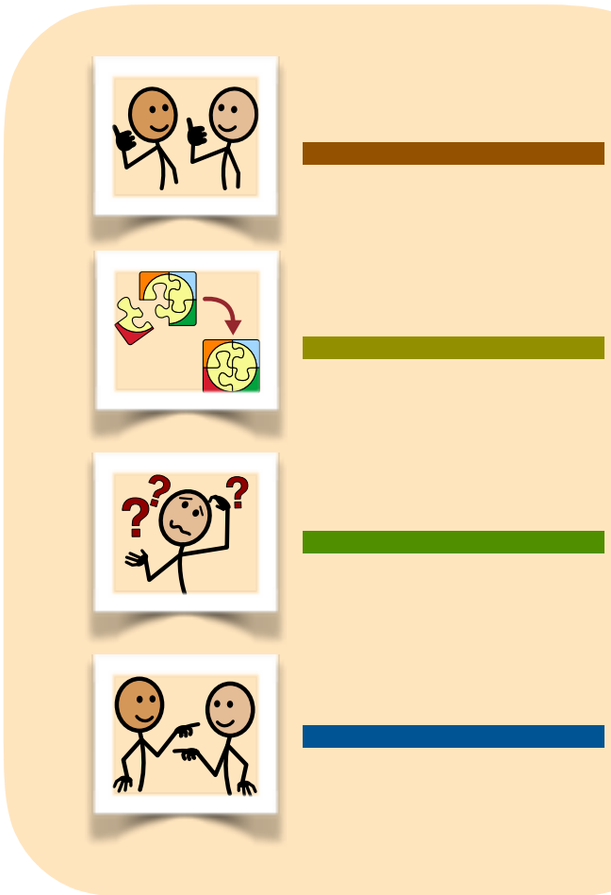


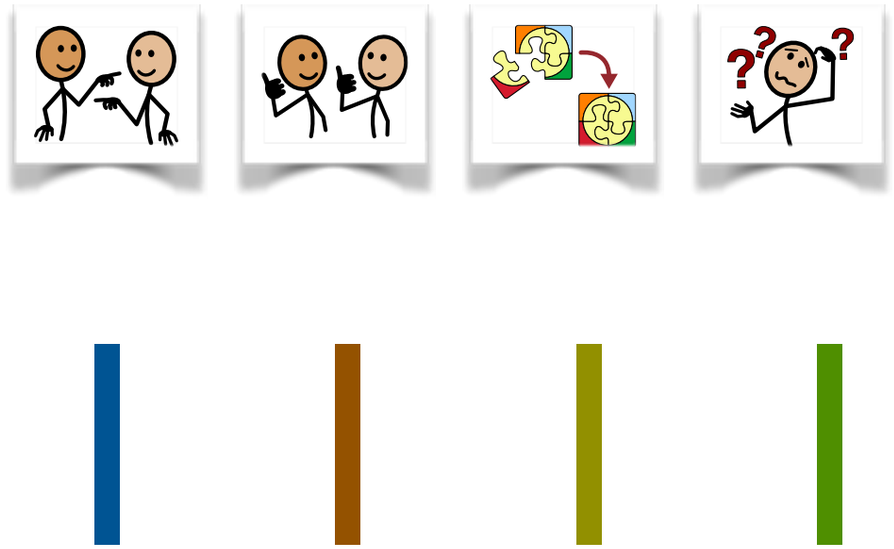






… you agree to solve **problems** …





Main Limitation or advantage:

Simulating Icon language through textual language

Future work:

Subjective evaluation by end users

Multi phrase representation for the icons

Multi sense representation for the icons

Additional material

The research paper: Compositional Language Modeling for Icon-Based Augmentative and Alternative Communication, DeepLo, ACL, 2018

The git repo: https://github.com/shiranD/icon_lm

Acknowledgements:

OHSU team:

Melanie Fried-Oken
Betts Peters
Brandon Eddie

Northeastern University team:

David Smith
Shaobin Xu

Our Funders:

My advisor: Steven Bedrick

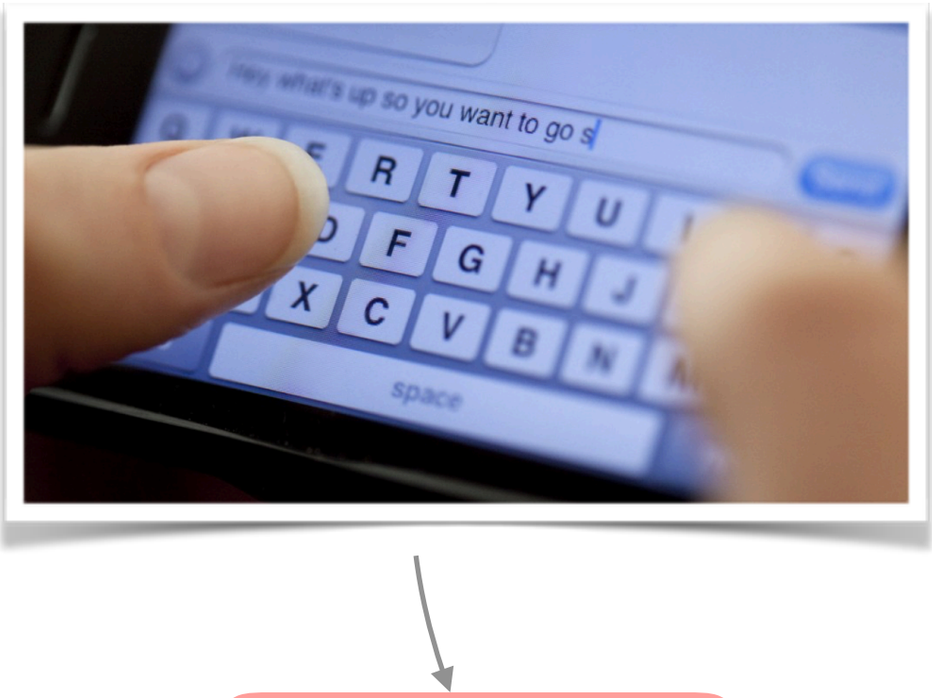
NIH, award
number 5R01DC009834-09

Toward Long Tailed LM with Continuous Output Prediction

Personalized Text Entry Use Case

FST-based

Personalized Text Entry Use Case



FST-based

Adapt to its user

Personalized Text Entry Use Case

FST-based

Adapt to its user

Readjust weights

Personalized Text Entry Use Case

FST-based

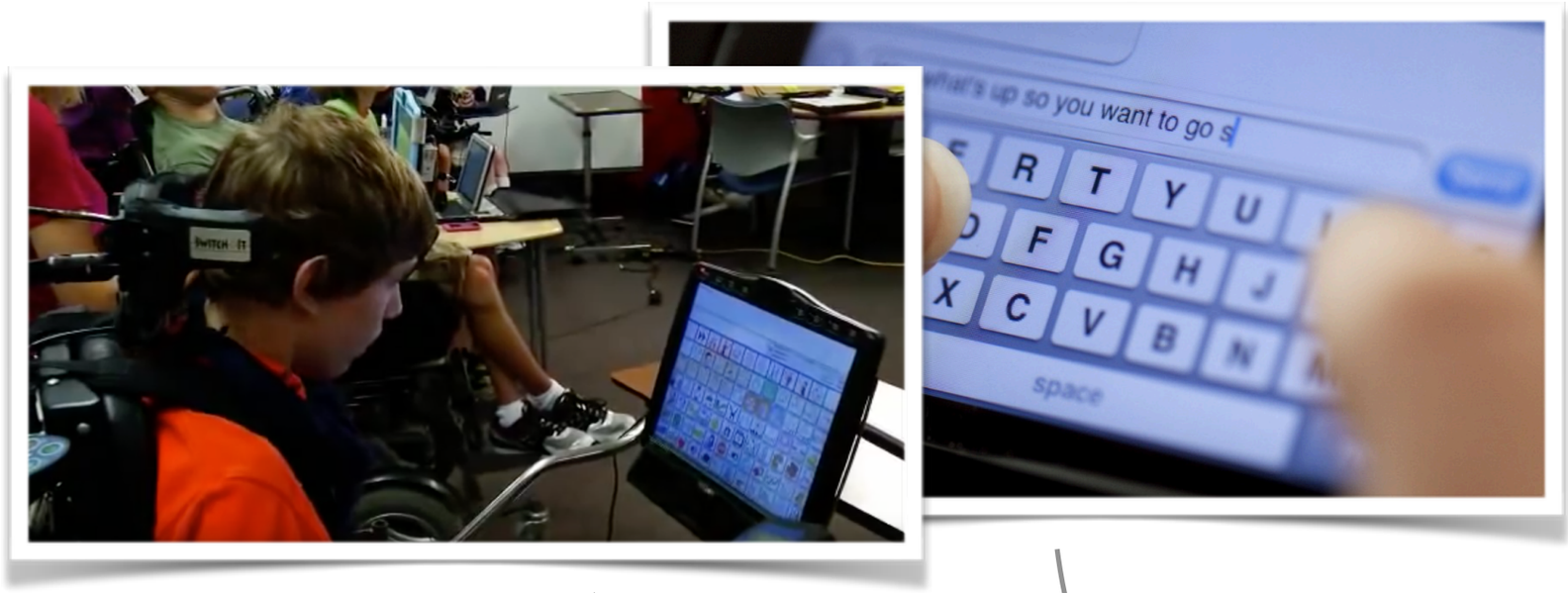
Adapt to its user

Readjust weights

Predict low frequency terms (personalization)

Add new vocabulary items

Personalized Text Entry Use Case



FST-based

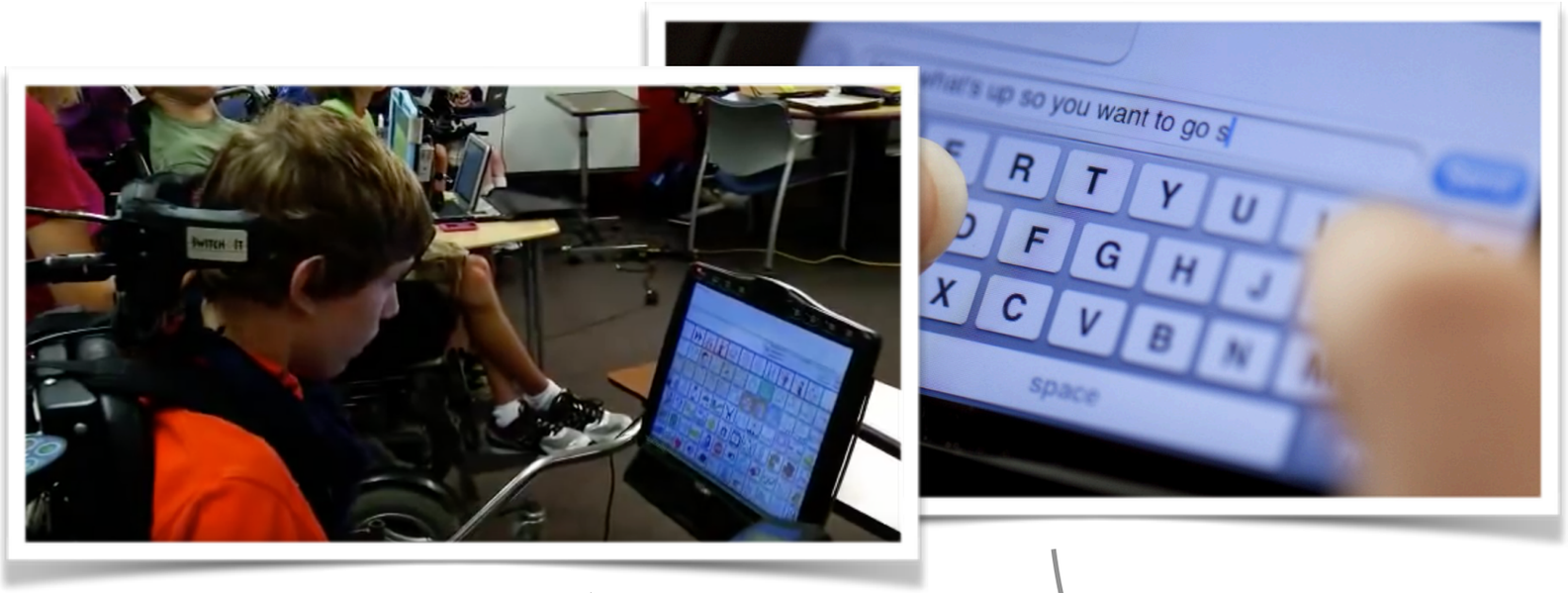
Adapt to its user

Readjust weights

Predict low frequency terms (personalization)

Add new vocabulary items

Personalized Text Entry Use Case



FST-based

Adapt to its user

Readjust weights

Predict low frequency terms (personalization)

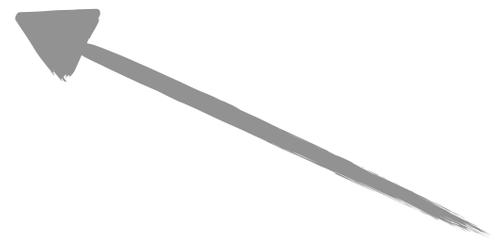
Add new vocabulary items

can

SotAs are softmax based

cannot

Fixed size vocabulary

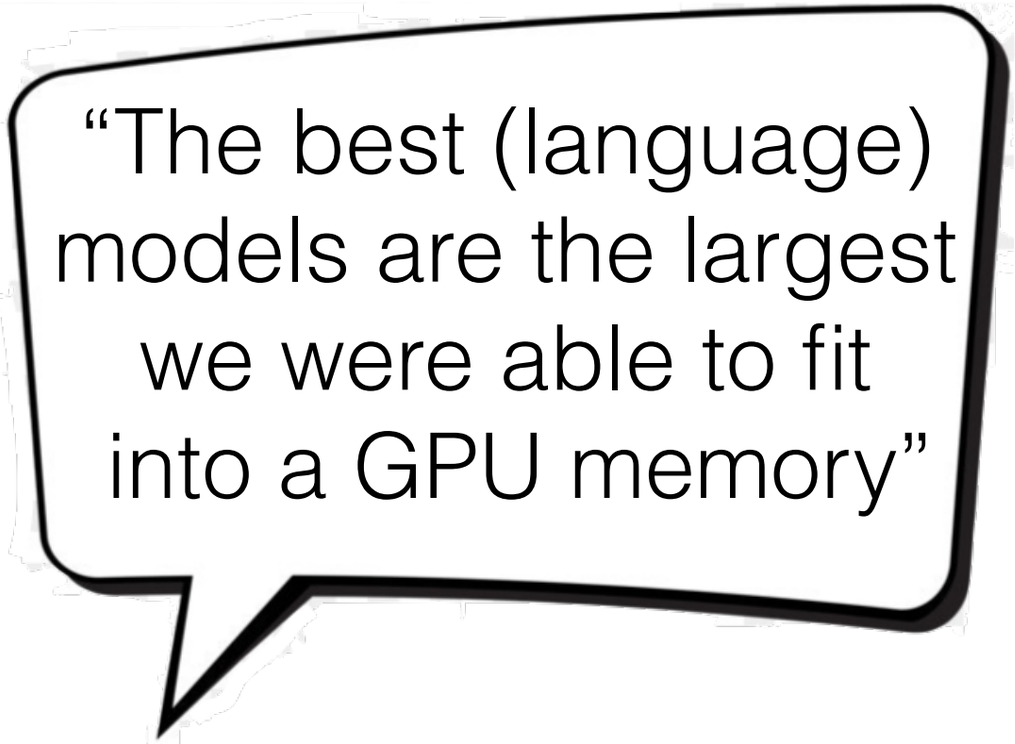


Problems with current approach

Fixed size vocabulary

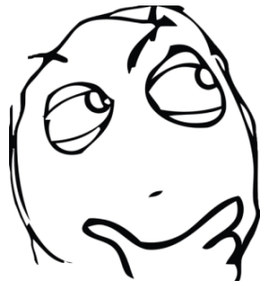
What else?

Problems with current approach



Josefowicz at el., 2016

why?

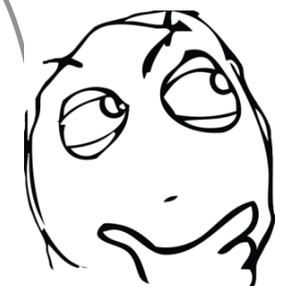


“The best (language) models are the largest we were able to fit into a GPU memory”

Josefowicz at el., 2016

Often, the deeper it gets the more parameters are required

why?

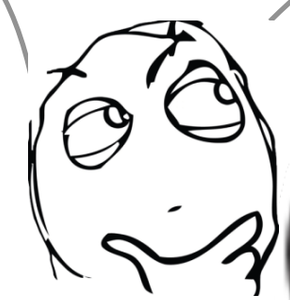


“The best (language) models are the largest we were able to fit into a GPU memory”

Josefowicz at el., 2016

Often, the deeper it gets the more parameters are required

why?



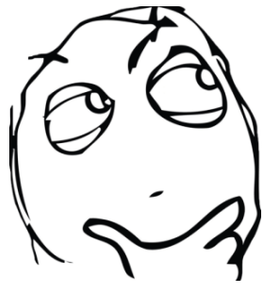
Current SotA models employ <u>categorical</u> prediction for Language Modeling which makes it, often, computationally intense (*)

"The best (language) models are the largest we were able to fit into a GPU memory"

Josefowicz at el., 2016

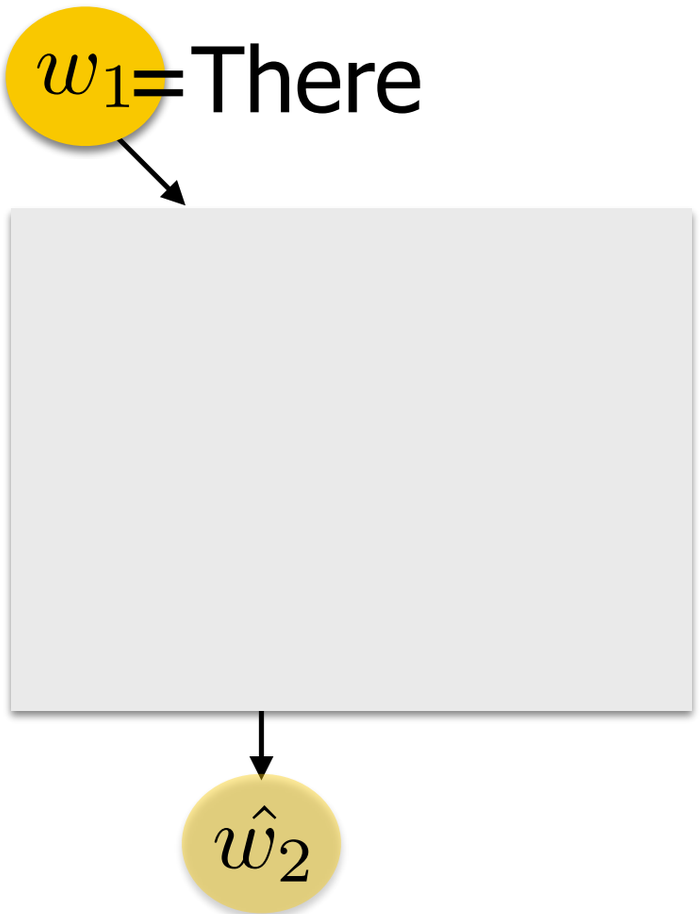
Current SotA models employ categorical prediction for Language Modeling which makes it, often, computationally intense (*)

wait,
how?



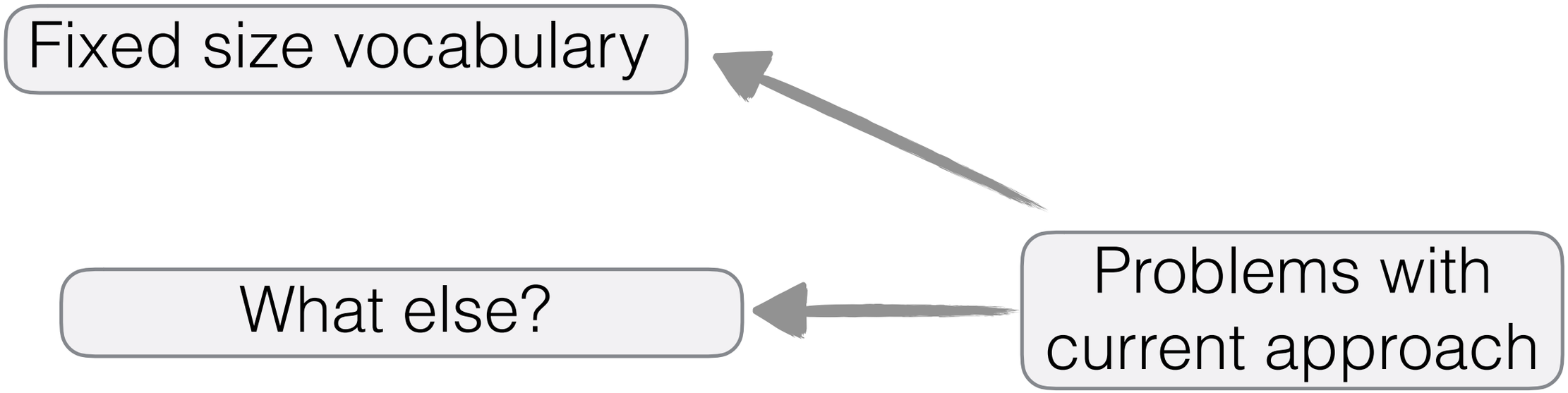
Current SotA models employ <u>categorical</u> prediction for Language Modeling which makes it, often, computationally intense (*)

SotA basic model

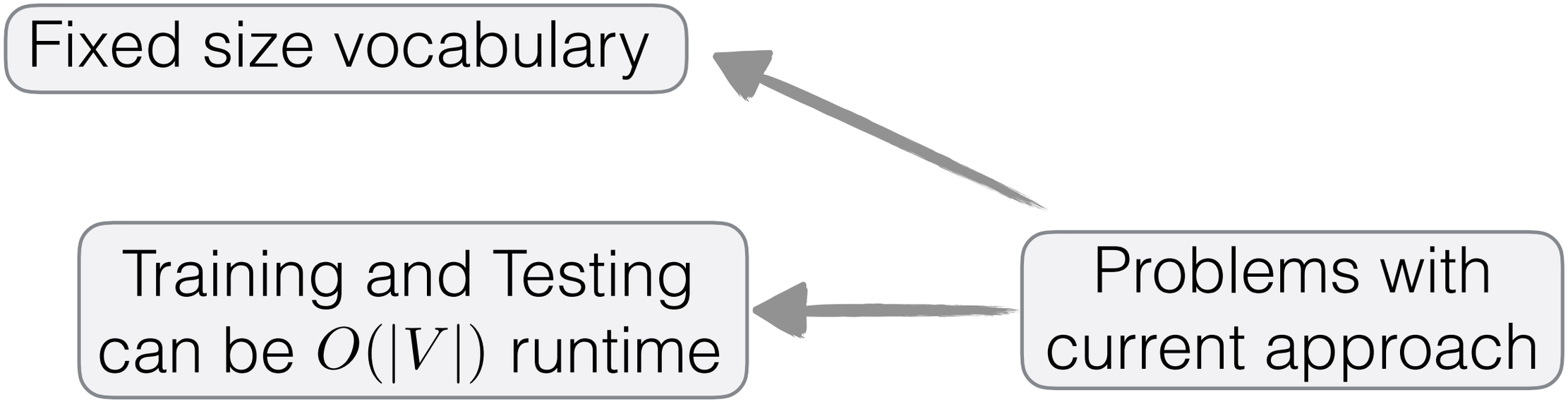


$$p(\hat{w}_i) = p(\hat{w}_i | w_{<i})$$

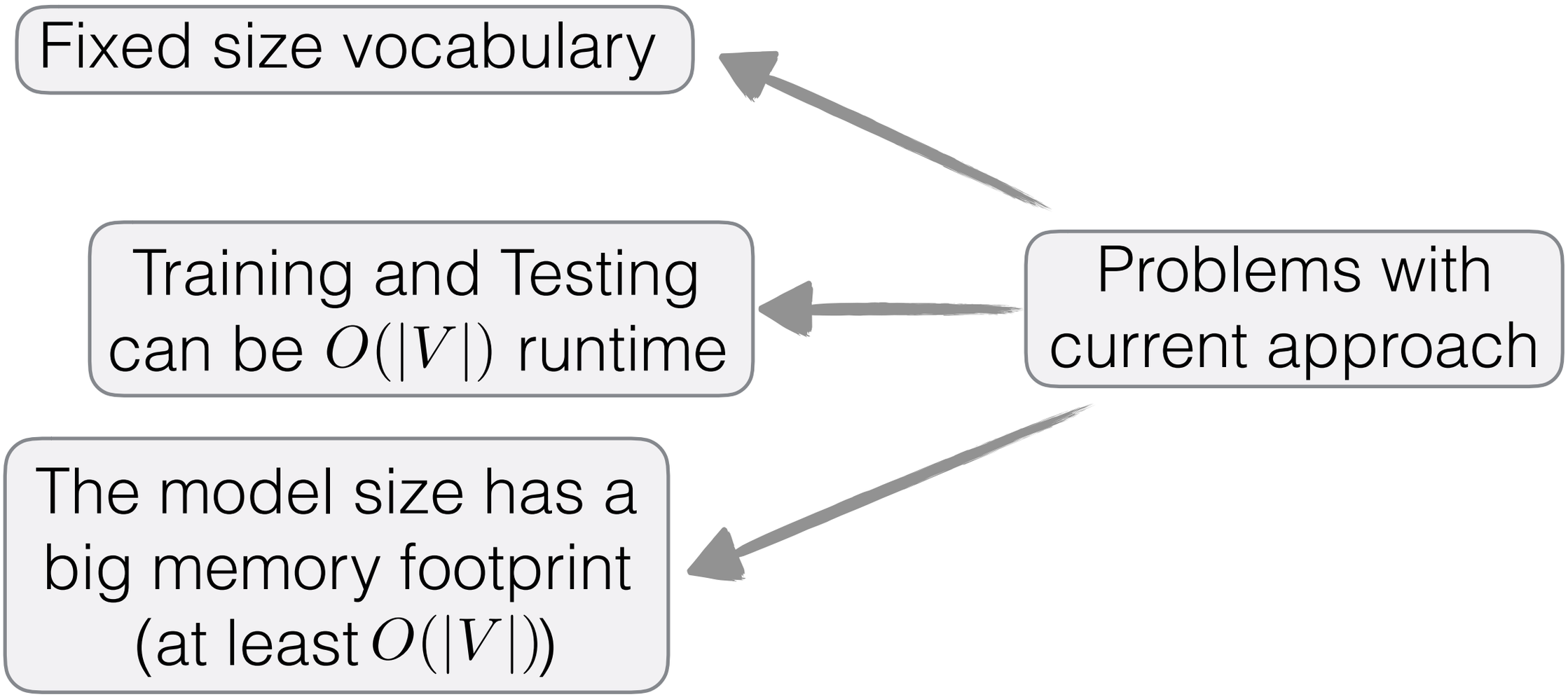
	likelihood (pr)
$\hat{w}_2 = are$	0.5
$\hat{w}_2 = is$	0.4
...	...
$\hat{w}_2 = up$	0.0007


Fixed size vocabulary
What else?
Problems with current approach

Model Complexities reflect mainly complexities of the final layer



Model Complexities reflect mainly complexities of the final layer



The model has
O(vocabulary)
memory requirement
just for its final layer

Josefowicz at el., 2016

Lets purely predict letters instead of words/subwords

overcome o(v) memory

The model has **O(vocabulary)** memory requirement just for its final layer

Josefowicz at el., 2016

Lets purely predict letters instead of words/subwords

but

Josefowicz at el., 2016

was worse than the other models

overcome o(v) memory

The model has **O(vocabulary)** memory requirement just for its final layer

Josefowicz at el., 2016

Lets purely predict letters instead of words/subwords

but

Josefowicz at el., 2016

was worse than the other models

besides

Dudy and Bedrick, 2018

Icon Language

Not all models can resort to such a small number of characters

overcome o(v) memory

The model has **O(vocabulary)** memory requirement just for its final layer

Josefowicz at el., 2016

Lets purely predict letters instead of words/subwords

but

Josefowicz at el., 2016

was worse than the other models

besides

Dudy and Bedrick, 2018

Icon Language

overcome o(v) memory

moreover

Not all models can resort to such a small number of characters

The model has **O(vocabulary)** memory requirement just for its final layer

``The ability of Bert to understand words depends highly on their frequency”

Schick and Schuze, 2019b

Josefowicz at el., 2016

Lets purely predict letters instead of words/subwords

but

Josefowicz at el., 2016

was worse than the other models

besides

Dudy and Bedrick, 2018

Icon Language

overcome o(v) memory

moreover

Not all models can resort to such a small number of characters

The model has **O(vocabulary)** memory requirement just for its final layer

``The ability of Bert to understand words depends highly on their frequency”

Schick and Schuze, 2019b

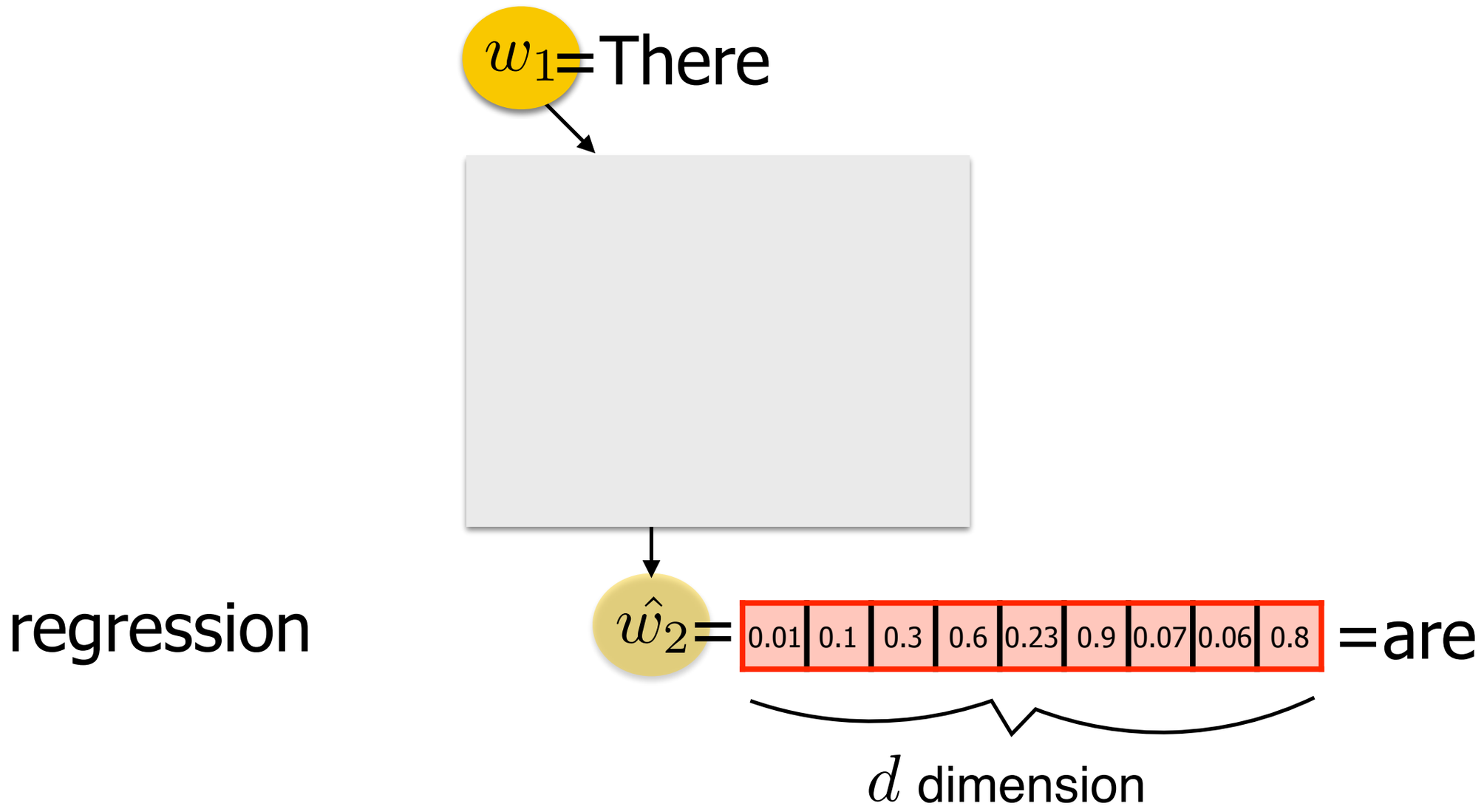
finally

Less appropriate for keyboard typing setting

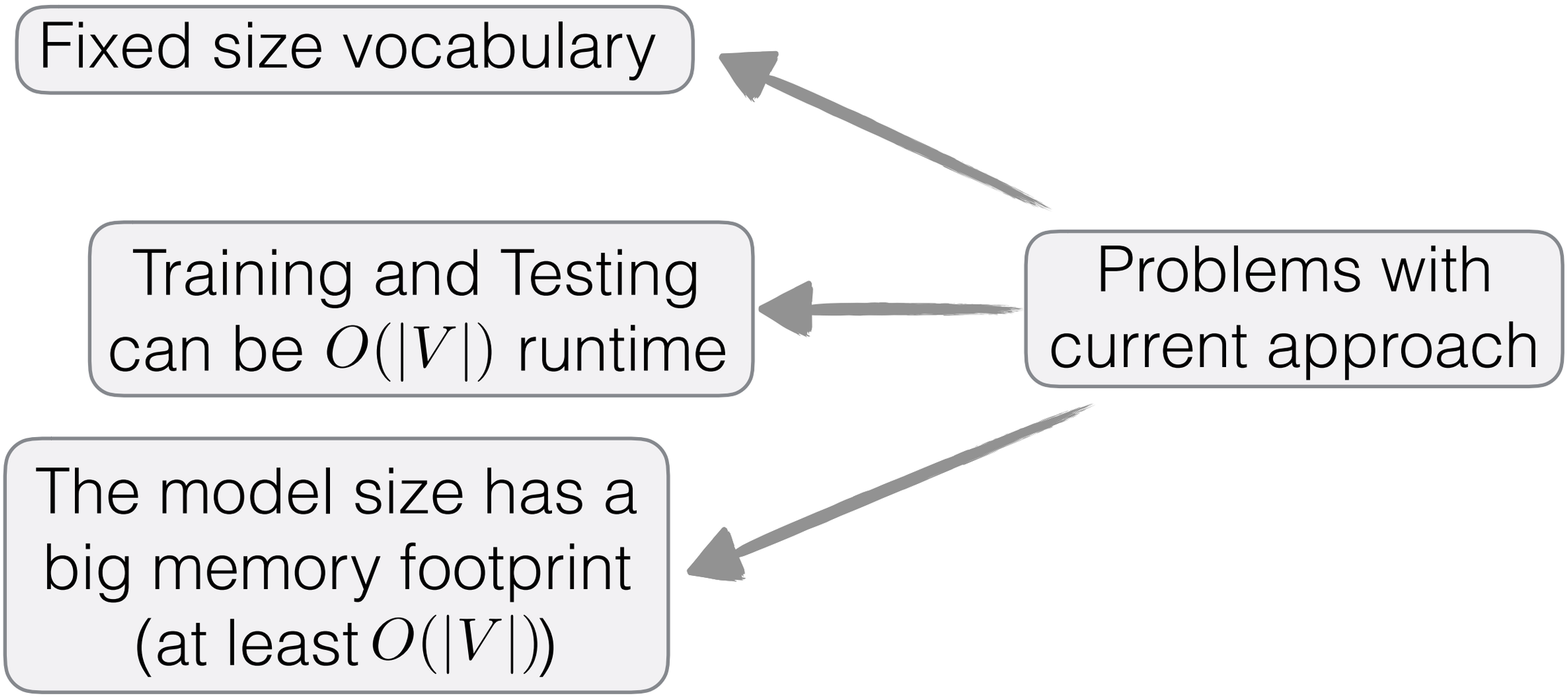
Our proposal

Train a language model to directly predict embeddings instead of discrete vocabulary entries and evaluate for low frequency words

Our Proposal



Model Complexities reflect mainly complexities of the final layer



Model Complexities reflect mainly complexities of the final layer

Continuous output prediction model

Fixed size vocabulary

Training and Testing can be $O(|V|)$ runtime

The model size has a big memory footprint (at least $O(|V|)$)

assumption: $d << |V|$

Model Complexities reflect mainly complexities of the final layer

potentially infinite

Fixed size vocabulary

Continuous output prediction model

Training and Testing can be $O(|V|)$ runtime

The model size has a big memory footprint (at least $O(|V|)$)

assumption: $d << |V|$

Model Complexities reflect mainly complexities of the final layer

potentially infinite

Fixed size vocabulary

$O(d)$

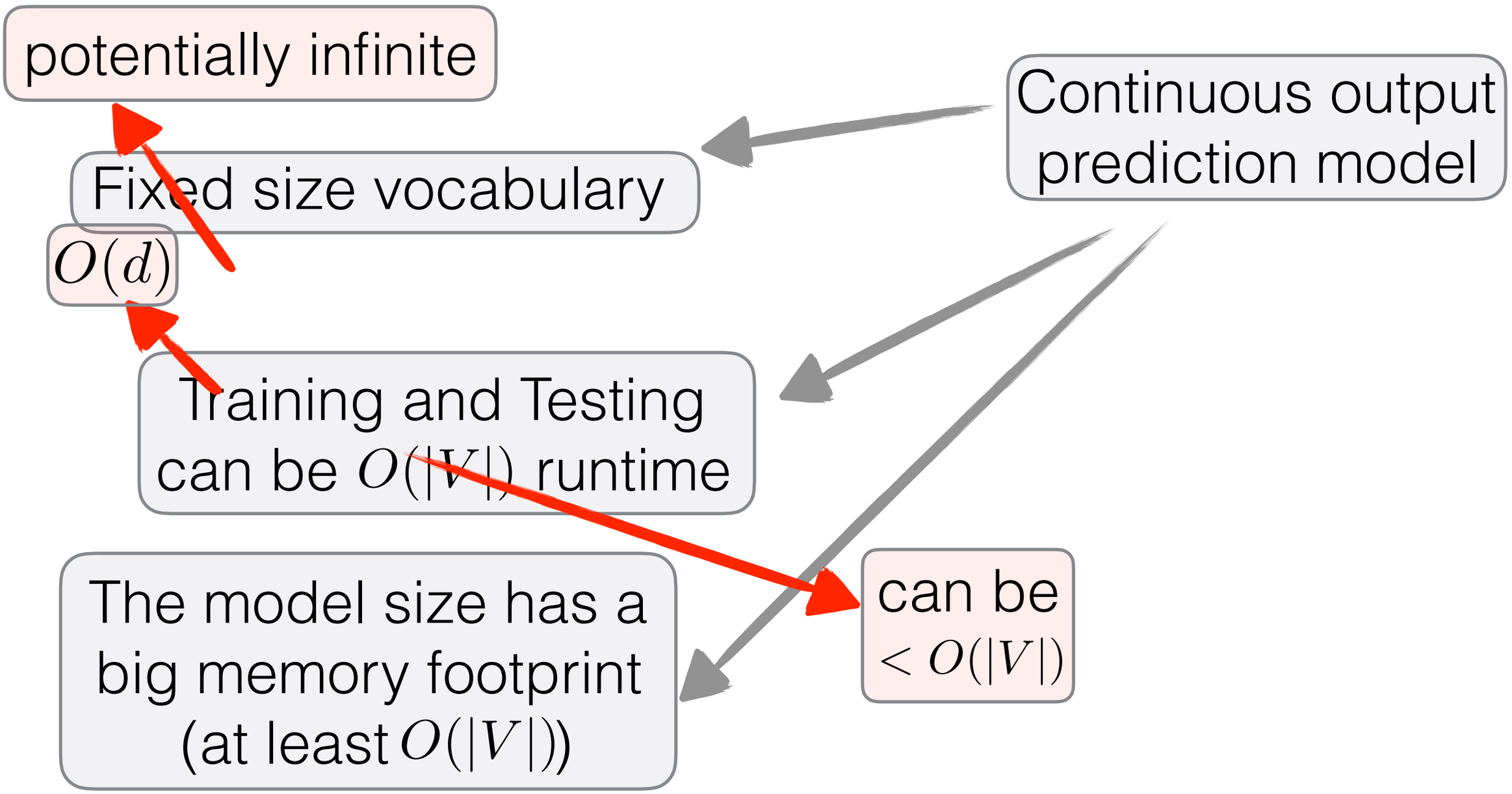
Training and Testing can be $O(|V|)$ runtime

Continuous output prediction model

The model size has a big memory footprint (at least $O(|V|)$)

assumption: $d << |V|$

Model Complexities reflect mainly complexities of the final layer



assumption: $d << |V|$

Model Complexities reflect mainly complexities of the final layer

potentially infinite

Fixed size vocabulary

$O(d)$

Training and Testing can be $O(|V|)$ runtime

Continuous output prediction model

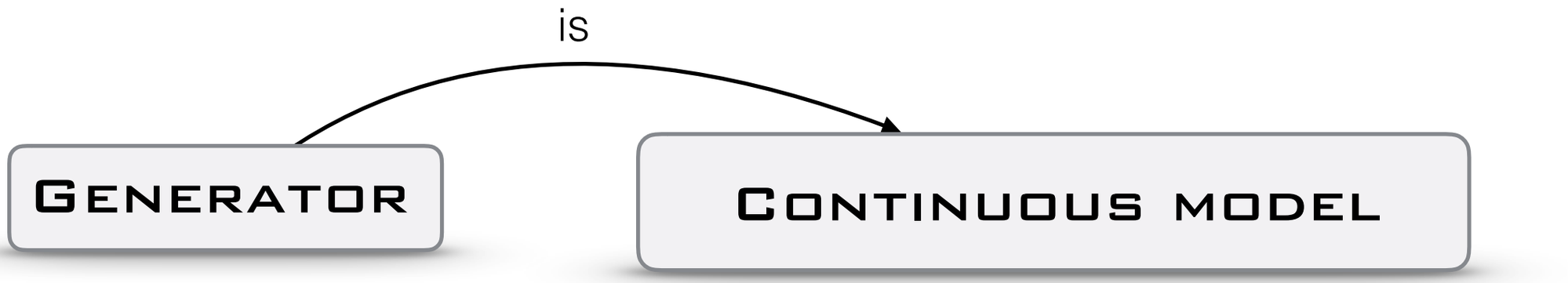
can be $< O(|V|)$

The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$

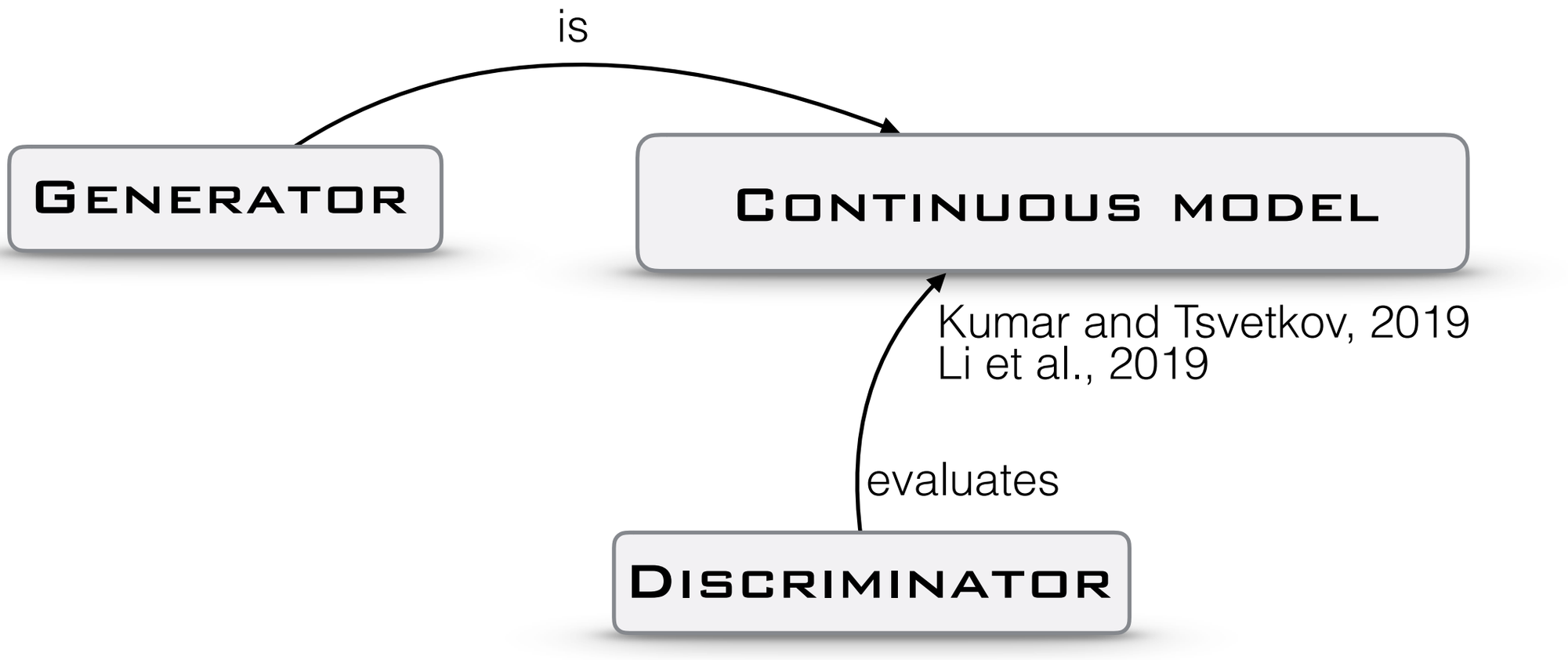
GAN Approach



Kumar and Tsvetkov, 2019
Li et al., 2019

models

GAN Approach



Data

Ferraro et al., 2018

Beck et al., 2010

experiment	num. types	num. tokens
NYT	950k	734M
PMC	860k	458M

Table 1: overall types and tokens

Decoding

Simple decoding

Nearest neighbor

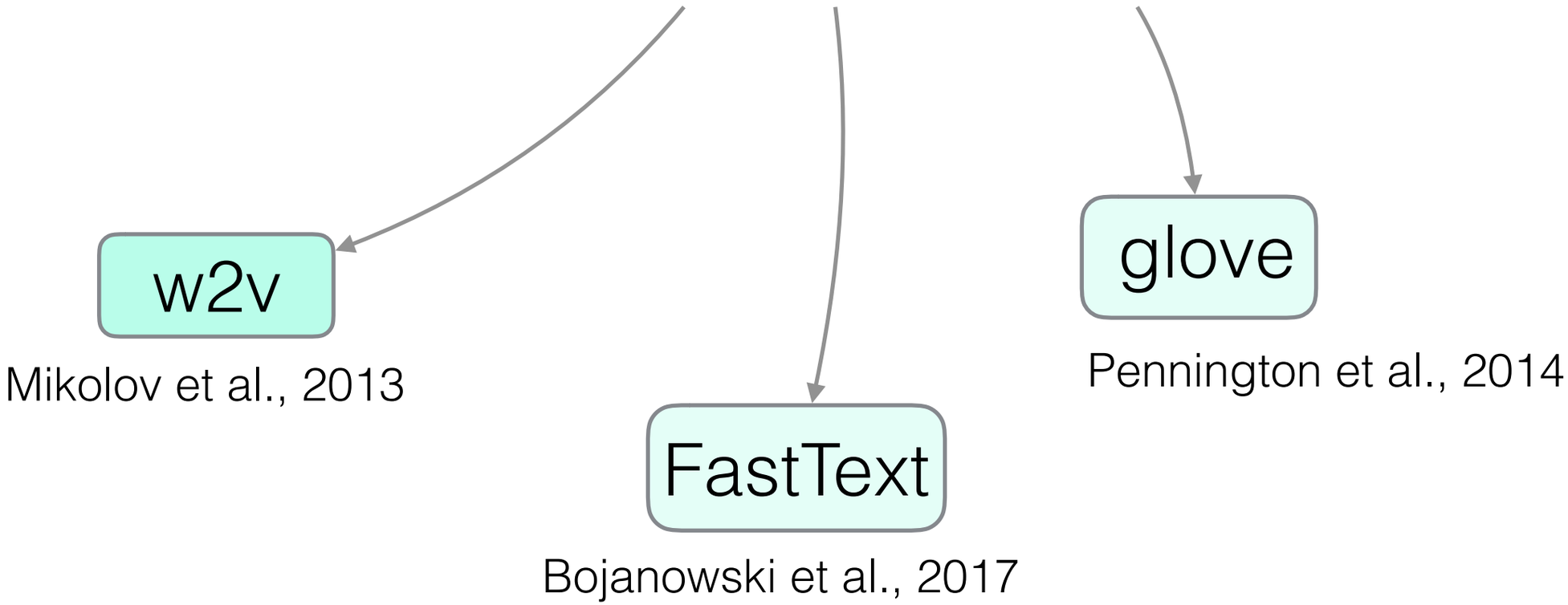
Feature-based decoding

PoS nearest neighbor

Ferraro et al., 2018

Neumann et al., 2019

Embedding Technique



Task

Predict the next term given a history of terms

Metrics

$top_1 (top_{10})$

percentage of trials that constituted a “hit”

Metrics

$top_1 (top_{10})$

percentage of trials that constituted a “hit”

$T_1 (T_{10})$

number of correctly predicted unique types

Metrics

$top_1(top_{10})$	percentage of trials that constituted a “hit”
$T_1(T_{10})$	number of correctly predicted unique types
MRR	Mean Reciprocal Rank of the targets

model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`			
`ugrm`			

Table 2: Experimental results on NYT corpus

model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`	00.89 (23.39)	1 (10)	0.04
`ugrm`			

Table 2: Experimental results on NYT corpus

model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`	00.89 (23.39)	1 (10)	0.04
`ugrm`	00.71 (08.46)	2,190 (5,592)	0.02

Table 2: Experimental results on NYT corpus

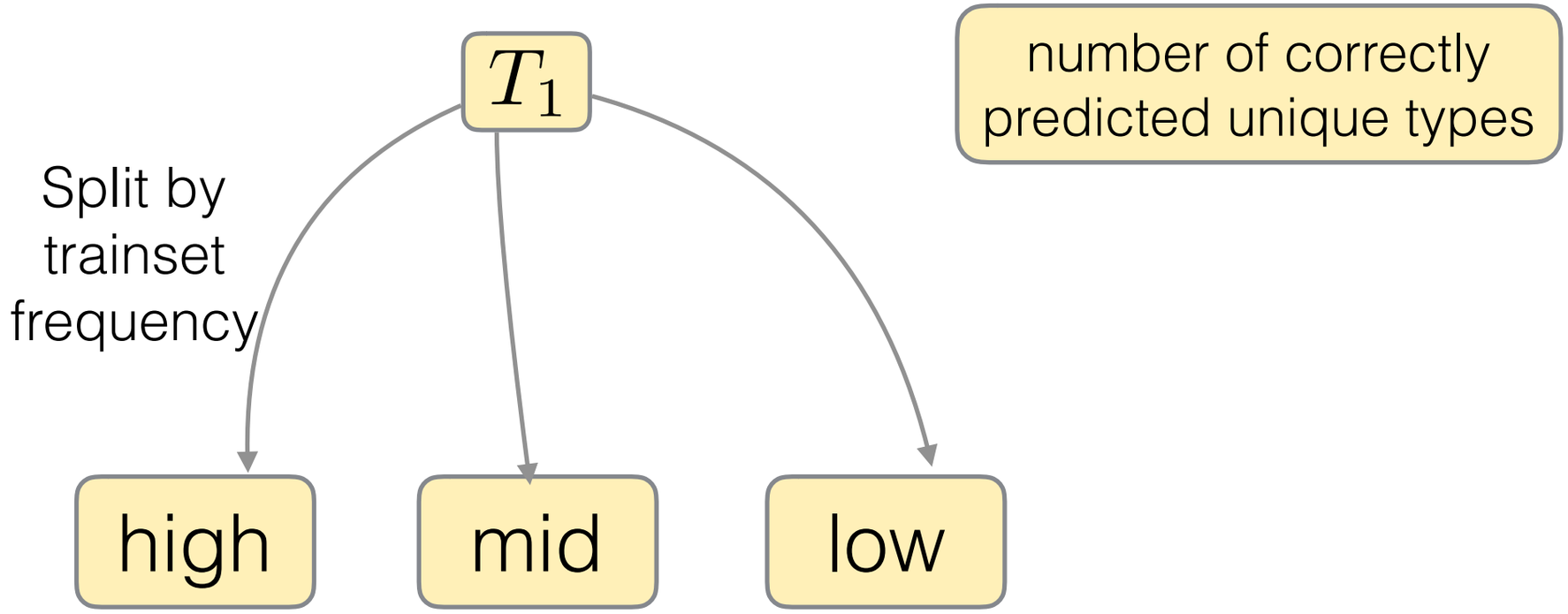
model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`	00.89 (23.39)	1 (10)	0.04
`ugrm`	00.71 (08.46)	2,190 (5,592)	0.02
`ctg`	19.19 (46.02)	3,982 (7,559)	0.27

Table 2: Experimental results on NYT corpus

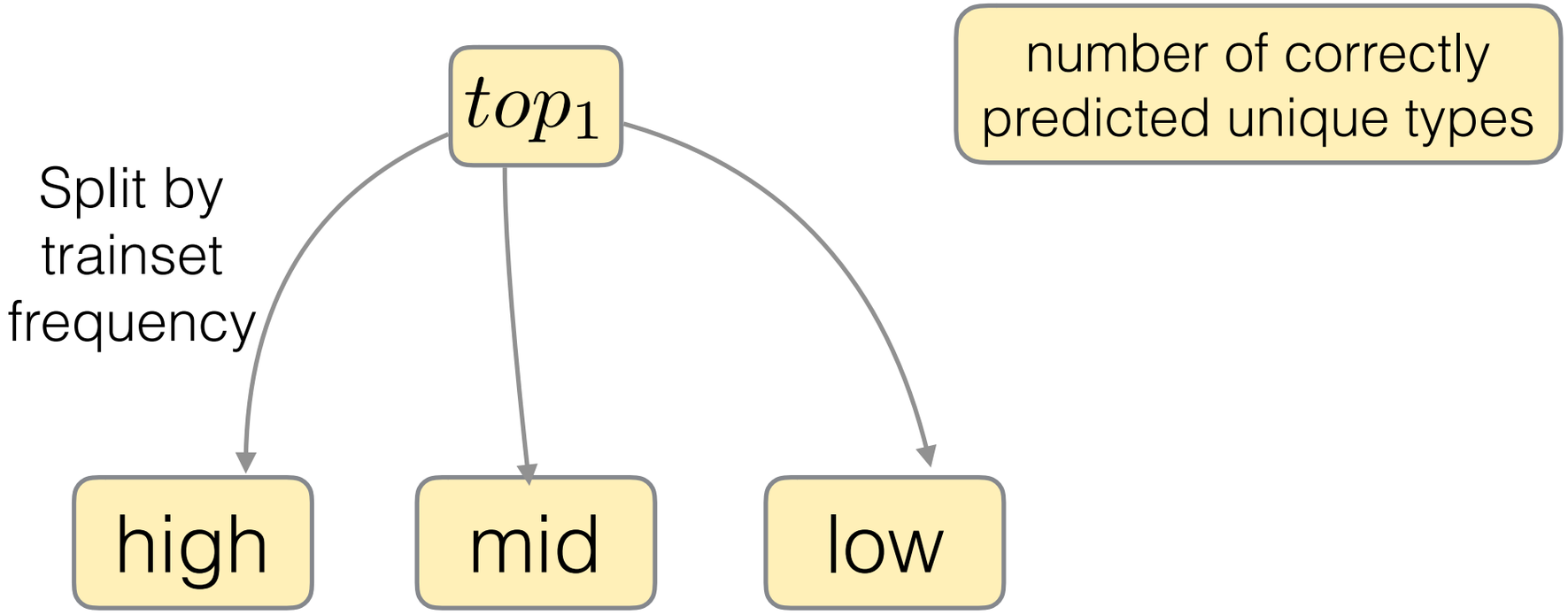
model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`	00.89 (23.39)	1 (10)	0.04
`ugrm`	00.71 (08.46)	2,190 (5,592)	0.02
`ctg`	19.19 (46.02)	3,982 (7,559)	0.27
$\mathsf{C}50$	17.31 (28.70)	8,917 (22,509)	0.21
$\mathsf{C}50_p$	20.09 (31.83)	10,326 (25,616)	0.24
$\mathsf{G}50$	16.46 (27.45)	11,534 (27,921)	0.20
$\mathsf{G}50_p$	19.56 (30.30)	13,540 (32,140)	0.23

Table 2: Experimental results on NYT corpus

Metrics


T_1
number of correctly predicted unique types
Split by trainset frequency
high
mid
low

Metrics


top_1
number of correctly predicted unique types
Split by trainset frequency
high
mid
low

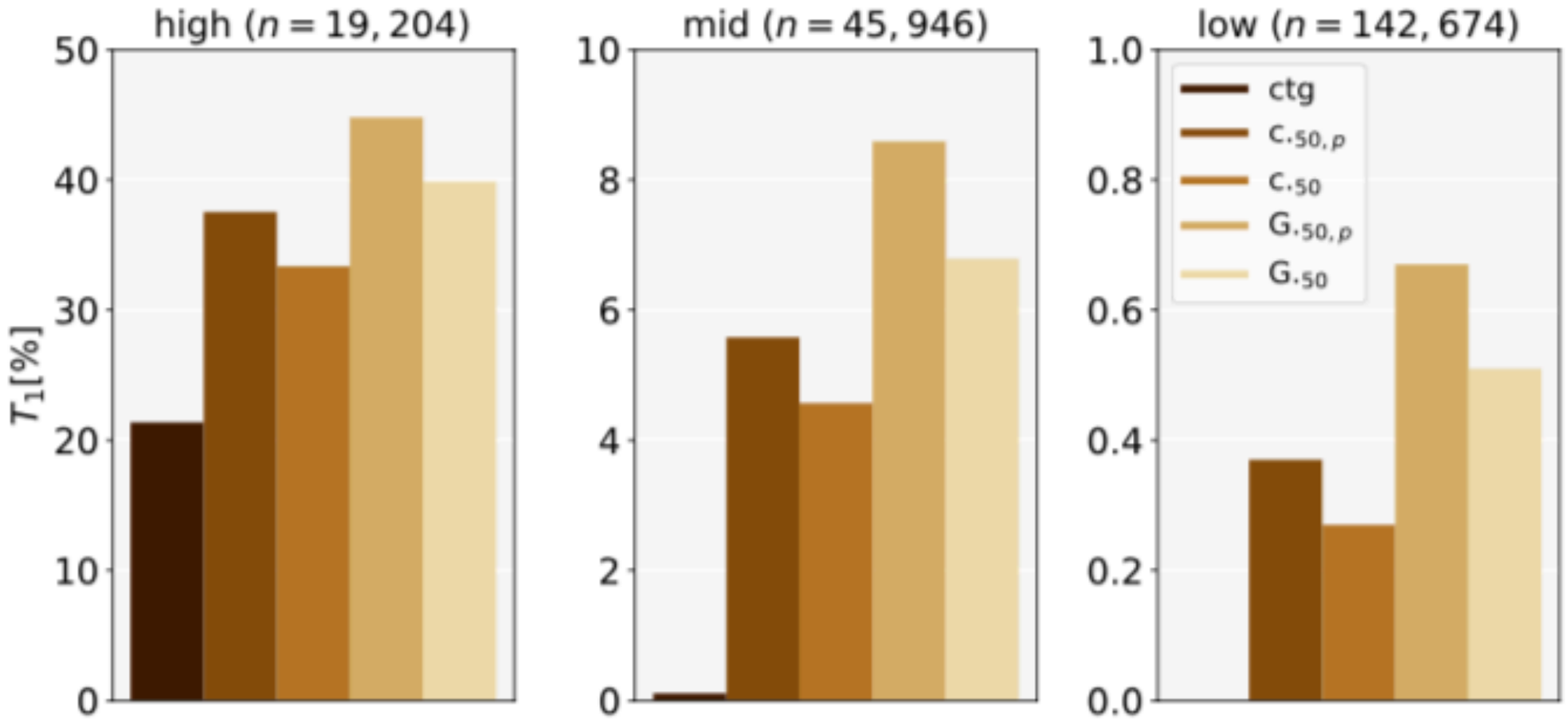


Figure 1: NYT *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

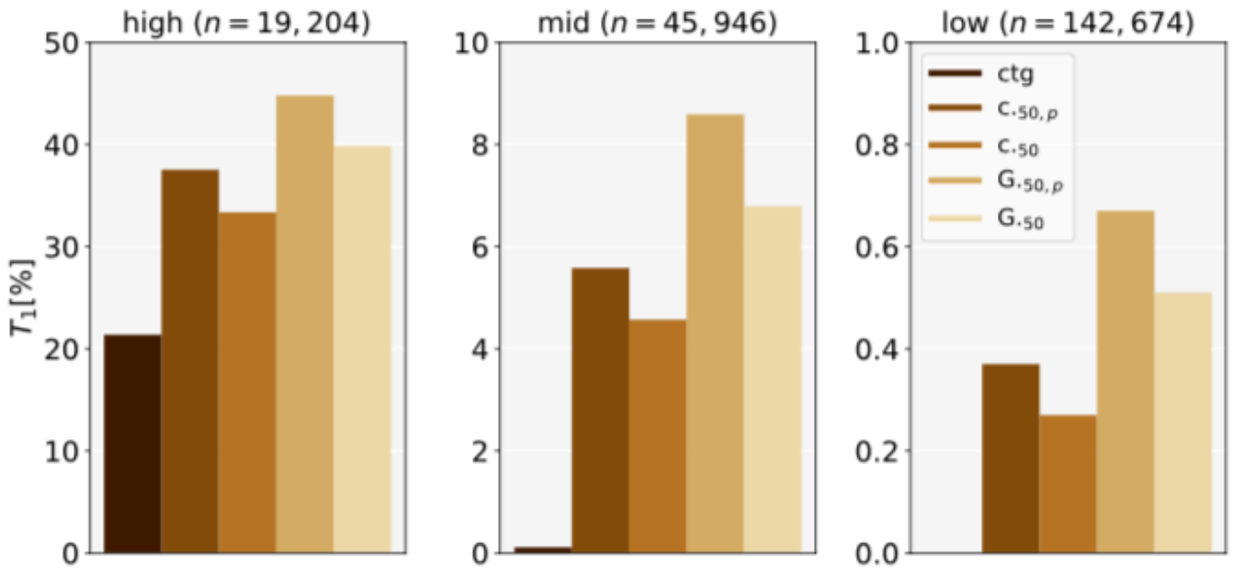


Figure 1: NYT *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

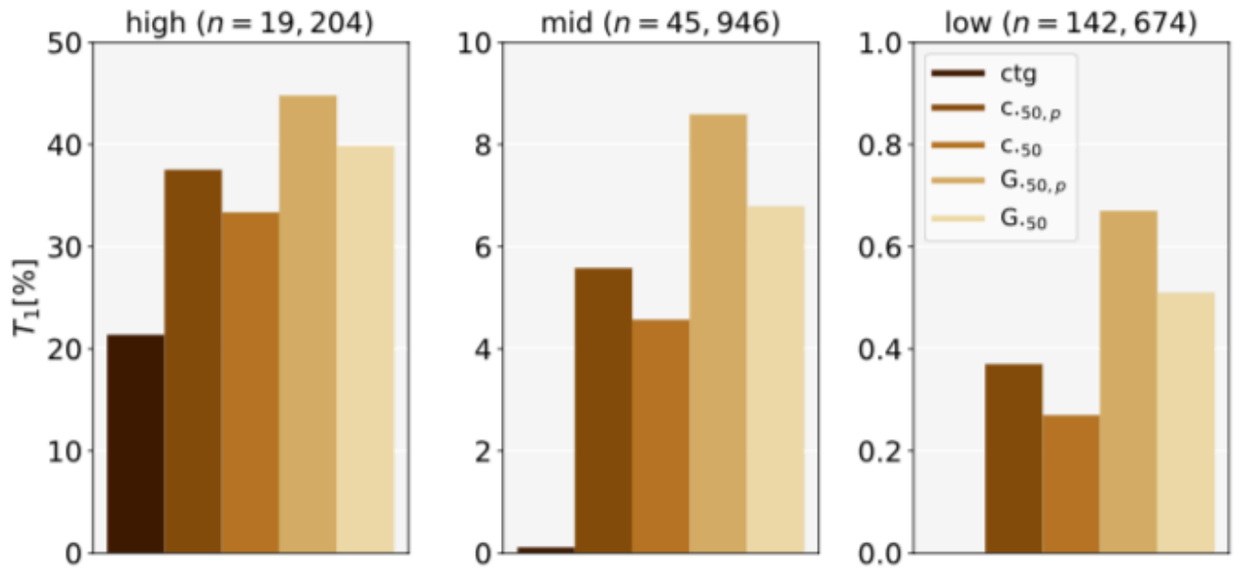


Figure 1: NYT *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

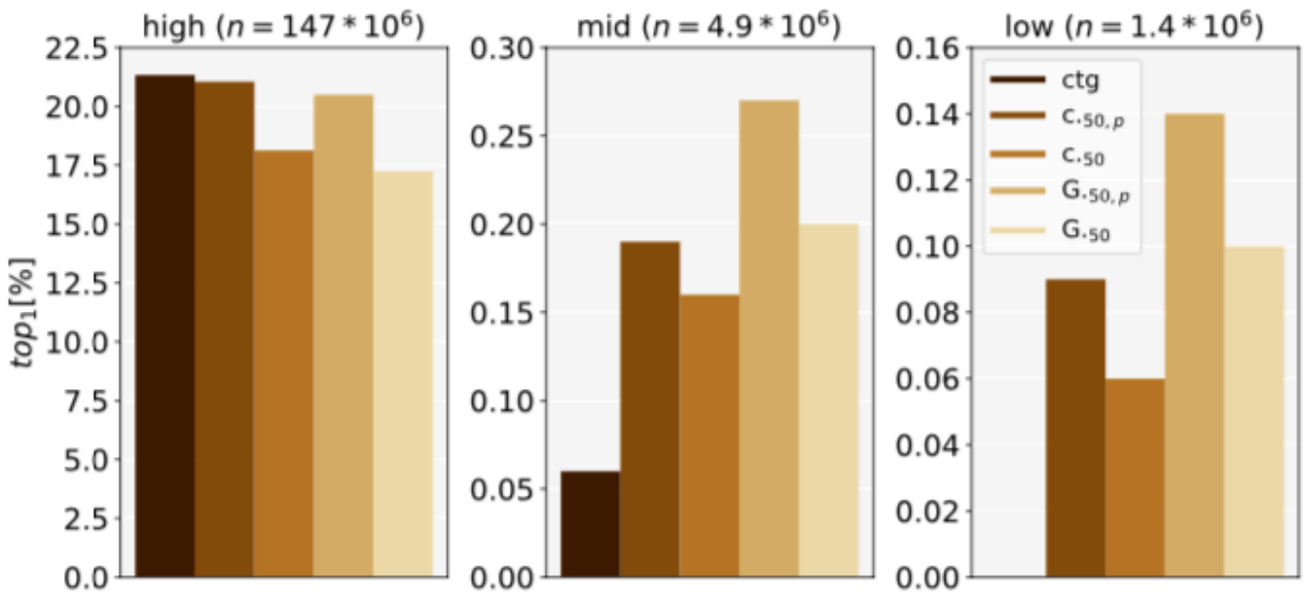


Figure 2: NYT *token* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

In this research

We showed how the continuous approach is more diverse* than a softmax based one on two different domains, NYT and PMC

We proposed a GAN based approach that provided relatively a richer type prediction

We proposed a feature based decoding mechanism that enhanced searches

Acknowledgements:

My advisor: Steven Bedrick

Questions?

shirdu2@gmail.com

model	top_1 (top_{10})	T_1 (T_{10})	M
`freq`	00.89 (25.53)	1 (10)	0.05
`ugrm`	00.76 (09.03)	1,790 (4,619)	0.02
`ctg`	22.12 (48.02)	4,764 (9,020)	0.30
$\mathsf{C50}$	19.89 (32.04)	11,947 (34,641)	0.24
$\mathsf{C50}_p$	22.35 (36.68)	15,909 (42,082)	0.27
$\mathsf{G50}$	19.04 (30.41)	18,040 (47,550)	0.23
$\mathsf{G50}_p$	21.32 (34.73)	23,544 (56,059)	0.26

Table 3: Experimental results on large PMC corpus

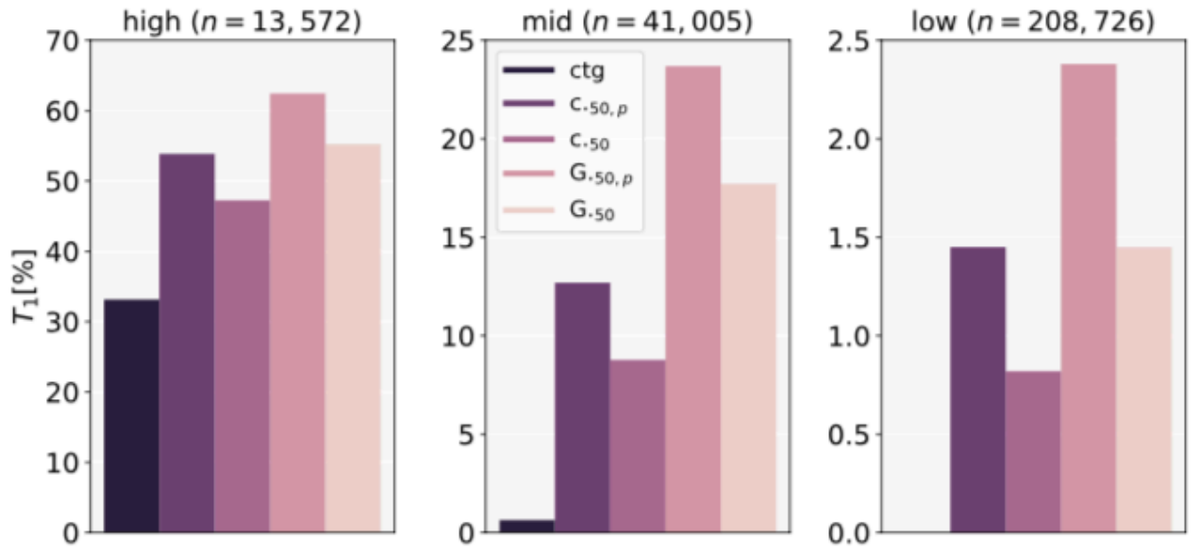


Figure 3: PMC *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

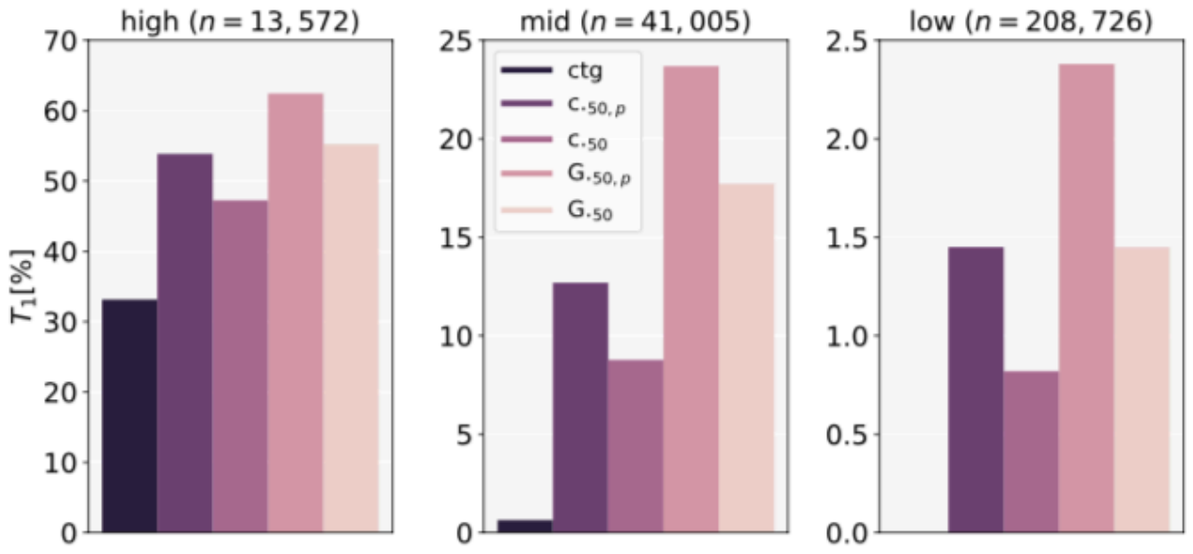


Figure 3: PMC *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

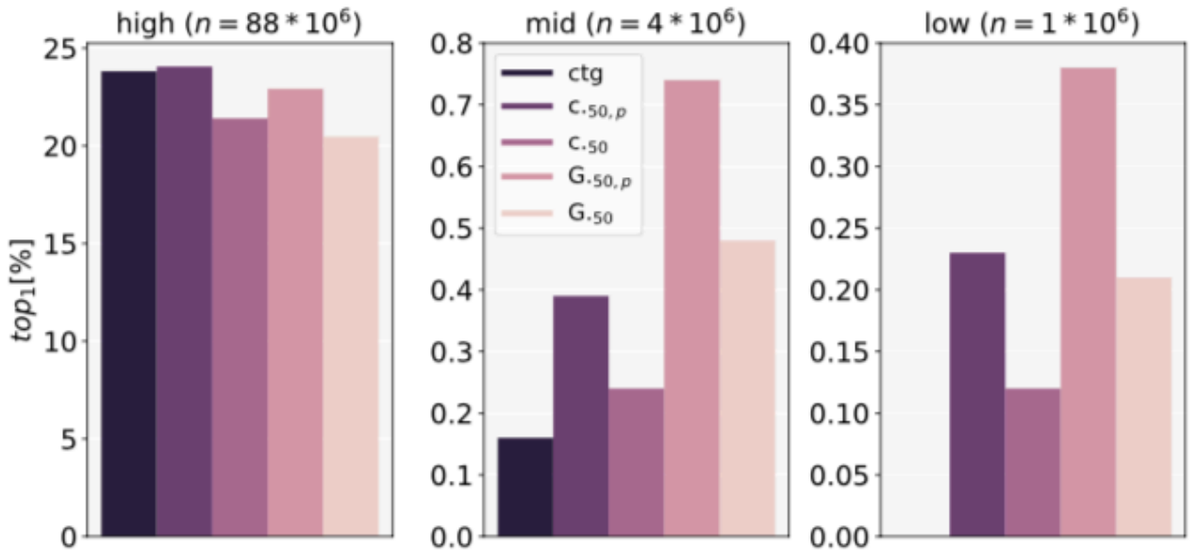


Figure 4: PMC *token* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

low bin	NYT predictions
G_p	the 34th floor is still hushed a privileged cocoon of mahogany paneling brass wall *sconces* ...
ctg	the 34th floor is still hushed a privileged cocoon of mahogany paneling brass wall *street* ...
G	a report on japan trade was issued today by the labor-industry coalition on international trade a coalition that brings together the afl-cio and major u.s. steel textile and semiconductor workers along with the economic strategy institute an *industry-financed* research group ...
ctg	a report on japan trade was issued today by the labor-industry coalition on international trade a coalition that brings together the afl-cio and major u.s. steel textile and semiconductor workers along with the economic strategy institute an *international* research group ...

low bin	PMC predictions
G	... peptides were eluted by centrifugation followed by NUM additional *elutions* with NUM μl
ctg	... peptides were eluted by centrifugation followed by NUM additional *NUM* with NUM μl
G_p	... are related to the constraint imposed on the electron *wavefunctions* by the surface
ctg	... are related to the constraint imposed on the electron *transport* by the surface
G	... owing to the anthropophilic and *endophilic* behaviour of ...
ctg	... owing to the anthropophilic and *unk* behaviour of ...
G_p	the abnormal karyotypes from of cytogenetic studies include autosomal trisomies sex chromosome monosomy triploidy double *trisomies* polyploidies ...
ctg	the abnormal karyotypes from of cytogenetic studies include autosomal trisomies sex chromosome monosomy triploidy double *and* polyploidies ...

mid bin	NYT predictions
G	my piano teacher at the paris *conservatory* a woman ...
ctg	my piano teacher at the paris *hotel* a woman ...
G	among the NUM american women who have hysterectomies each year thousands may die prematurely of heart disease because doctors removed their *ovaries* along with ...
ctg	among the NUM american women who have hysterectomies each year thousands may die prematurely of heart disease because doctors removed their *way* along with ...
G	this mold covers the leaf in soot preventing the cells underneath from absorbing sunlight and conducting *photosynthesis* eventually ...
ctg	this mold covers the leaf in soot preventing the cells underneath from absorbing sunlight and conducting *the* eventually ...
mid bin	**PMC predictions**
G	this is further confirmed by detection of apoptotic and *nonapoptotic* death ...
ctg	this is further confirmed by detection of apoptotic and *cell* death ...
G	in particular our objectives were to determine the association between NUM reported alcohol *misuse* and hiv sexual ...
ctg	in particular our objectives were to determine the association between NUM reported alcohol *and* and hiv sexual ...
G	human infections are common following handling or processing of infected turkeys or *ducks*
ctg	human infections are common following handling or processing of infected turkeys or *by*
G	NUM patients underwent cardiopulmonary resuscitation including NUM patients with respiratory arrest and pronounced *bradycardia*
ctg	NUM patients underwent cardiopulmonary resuscitation including NUM patients with respiratory arrest and pronounced *NUM*

Table 9: rare words' correct and incorrect predictions of continuous and categorical models respectively

mid bin	NYT predictions
ctg	-lrb- eyman palm beach post -rrb- art includes a *6-pica* color photo of book jacket
G	-lrb- eyman palm beach post -rrb- art includes a *compilation* color photo of book jacket
ctg	-lrb- this article is *excerpted* from new scientist ...
G	-lrb- this article is *included* from new scientist ...
ctg	the moscow military parade which has drawn no criticism across the russian political spectrum will put on display some of the very types of weapons that have been pounding chechnya since last december including t-72 tanks grad rocket-salvo *launchers* and NUM ...
G	the moscow military parade which has drawn no criticism across the russian political spectrum will put on display some of the very types of weapons that have been pounding chechnya since last december including t-72 tanks grad rocket-salvo *helicopters* and NUM ...

mid bin	PMC predictions
ctg	the sections were prepared by ultra microtone leica *microsystems* and stained ...
G	the sections were prepared by ultra microtone leica *kontron* and stained ...
ctg	height and weight was measured by trained technicians using standardized equipment with participants wearing light *clothing* without shoes
G	height and weight was measured by trained technicians using standardized equipment with participants wearing light *shoes* without shoes
ctg	this gave recommended retail prices rrp for all major cigarette brands for the uk market ie great *britain* and northern ireland
G	this gave recommended retail prices rrp for all major cigarette brands for the uk market ie great *livelihood* and northern ireland

Table 10: mid bin correct and incorrect predictions of categorical and continuous models respectively

potentially infinite

Fixed size vocabulary

$O(d)$

Training and Testing can be $O(|V|)$ runtime

Continuous output prediction model

can be $< O(|V|)$

The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$

Continuous output prediction model

The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$

decoding model

$O(|V|)$

user model

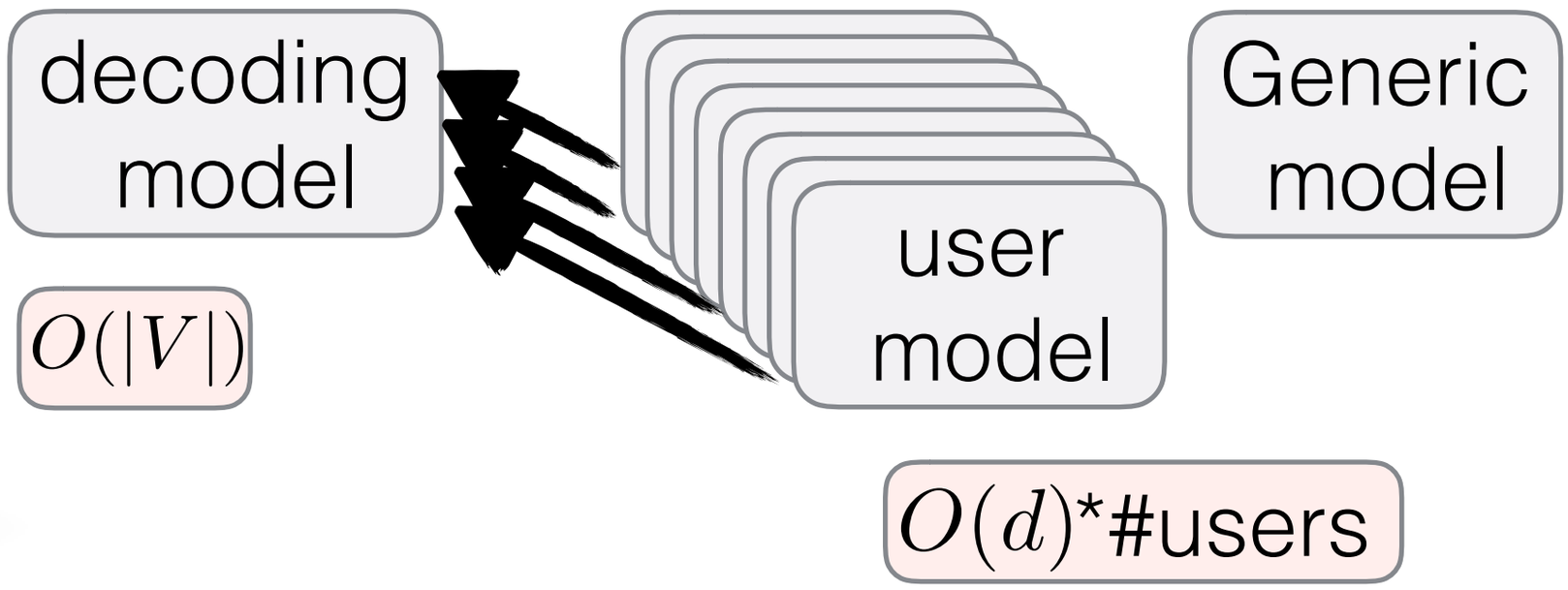
$O(d)$

Generic model

The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$



The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$

decoding model

$O(|V|)$

user model

Generic model

$O(d)$*#users

categorical approach: $O(|V|)$*#users

The model size has a big memory footprint (at least $O(|V|)$)

$O(d)$

assumption: $d << |V|$

The limited Bandwidth of EEG based BCIs

P300 waveform associated w consciousness

Passive Frame theory: A new synthesis. Behavioral and Brain Sciences, Ezequiel Morsella et al, 2016

"When you speak, you're only consciously aware of a few words at a time, and that is only so you can direct the muscles around your mouth and tongue to form those words. What you're saying is prescribed under the hood; your conscious mind is simply following a script."

What are Language Models?

Language Modeling Module

A statistical **language model** is a probability distribution over sequences of tokens (words is a special case)

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

Language Modeling Module

A statistical **language model** is a probability distribution over sequences of tokens (words is a special case)

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

$p(\text{it is simple}) = ?$

Language Modeling Module

A statistical **language model** is a probability distribution over sequences of tokens (words is a special case)

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

$p(\text{it is simple}) = ?$

$p(\text{it is simple}|unigram) = p(\text{it})p(\text{is})p(\text{simple})$

Language Modeling Module

A statistical **language model** is a probability distribution over sequences of tokens (words is a special case)

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

$$p(\text{it is simple}) = ?$$
$$p(\text{it is simple}|unigram) = p(\text{it})p(\text{is})p(\text{simple})$$
$$p(\text{it is simple}|bigram) = p(\text{it}|<s>)p(\text{is}|\text{it})p(\text{simple}|\text{is})p(</s>|\text{simple})$$

Language Modeling Module

A statistical **language model** is a probability distribution over sequences of tokens (words is a special case)

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

$$p(\text{it is simple}) = ?$$
$$p(\text{it is simple}|unigram) = p(\text{it})p(\text{is})p(\text{simple})$$
$$p(\text{it is simple}|bigram) = p(\text{it}|<s>)p(\text{is}|\text{it})p(\text{simple}|\text{is})p(</s>|\text{simple})$$

How to define probability distributions over strings?

We have embedding in sequences

We have embedding in sequences

Now we are ready for Icon LM training

We have embedding in sequences

Now we are ready for Icon LM training

Simple (any architecture we want)

We have embedding in sequences

Now we are ready for Icon LM training

Simple (any architecture we want)

5 fold cross validation

We have embedding in sequences

Now we are ready for Icon LM training

Simple (any architecture we want)

5 fold cross validation

LM prediction accuracy: MRR, ACC@1, ACC@10

We have embedding in sequences

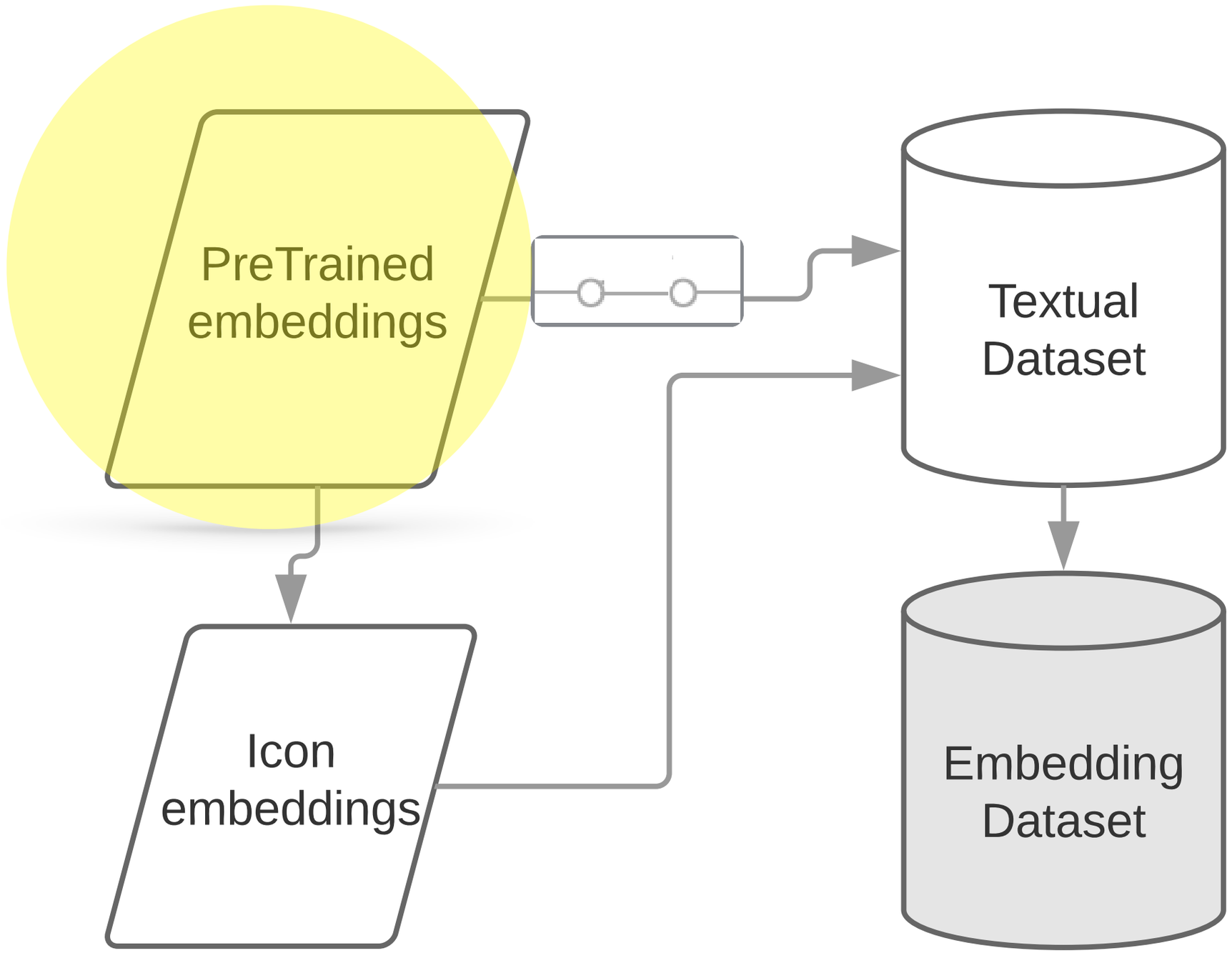
Now we are ready for Icon LM training

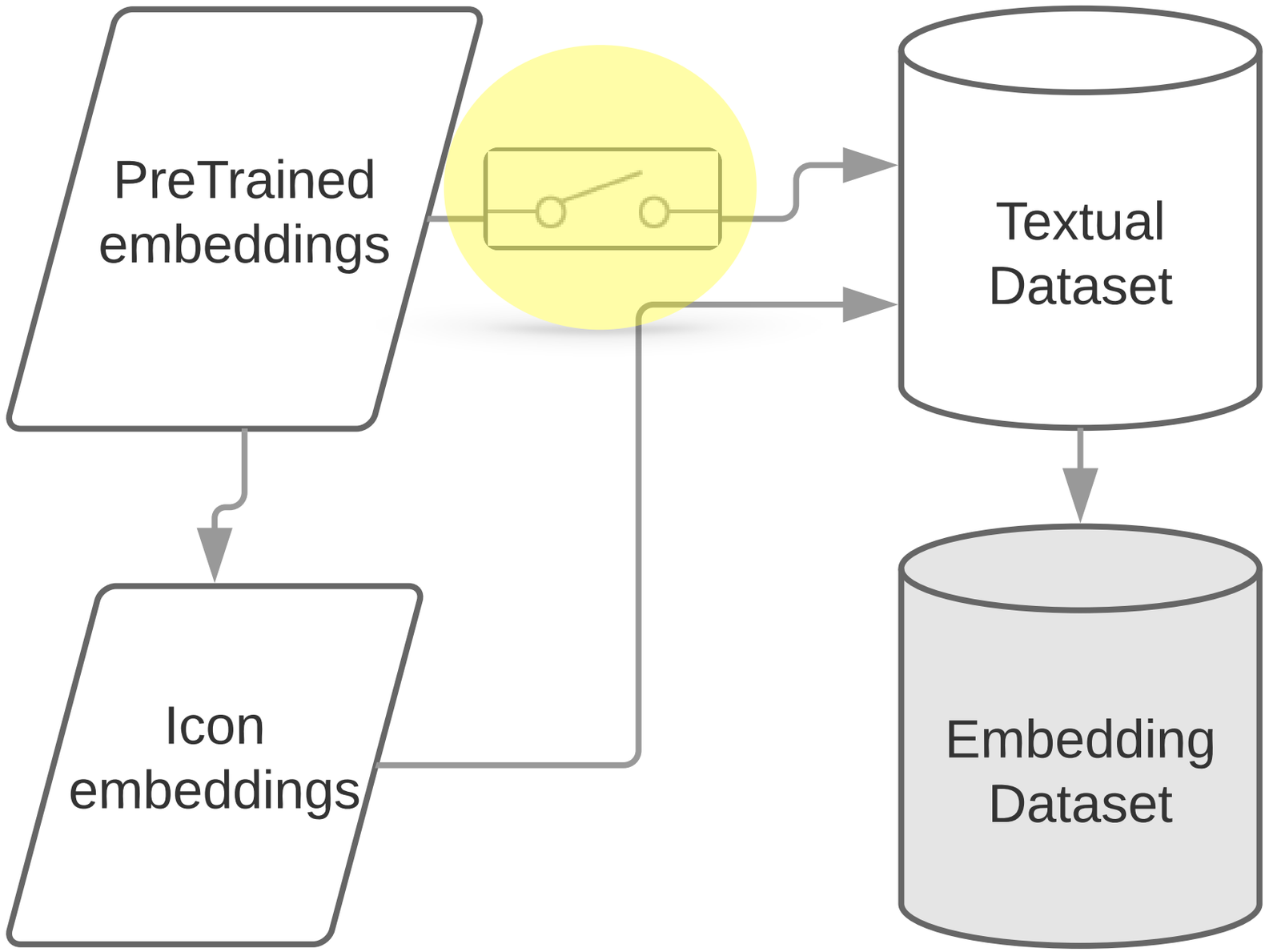
Simple (any architecture we want)

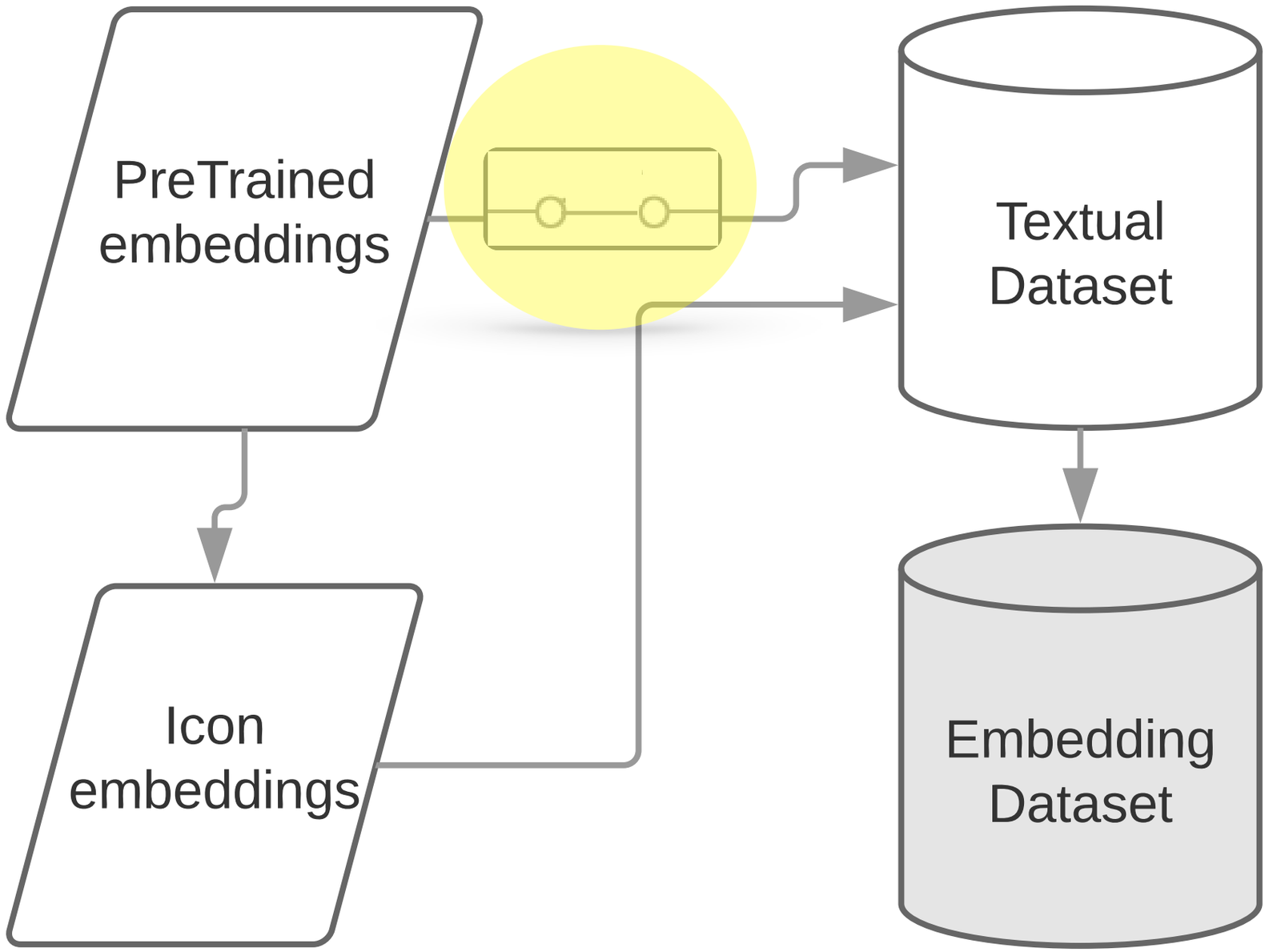
5 fold cross validation

LM prediction accuracy: MRR, ACC@1, ACC@10

How to evaluate our choices throughout the process


PreTrained
embeddings
Textual
Dataset
Icon
embeddings
Embedding
Dataset

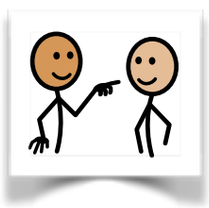

PreTrained
embeddings
Textual
Dataset
Icon
embeddings
Embedding
Dataset


PreTrained
embeddings
Textual
Dataset
Icon
embeddings
Embedding
Dataset

Experiment 2: Icon constraint

English: “your warning did not work”

non-pure:




<your> <warning> <did> <not> <work>

Experiment 2: Icon constraint

English: “your warning did not work”

non-pure:

<your> <warning> <did> <not> <work>

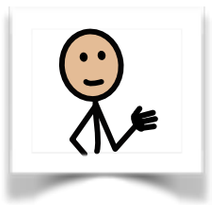
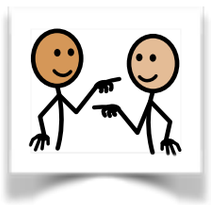
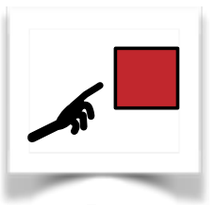
pure:

<your> <warning> <not> <work>

Experiment 2: Icon constraint

English: “so you did it”

non-pure:

<so> <you> <did> <it>

Experiment 2: Icon constraint

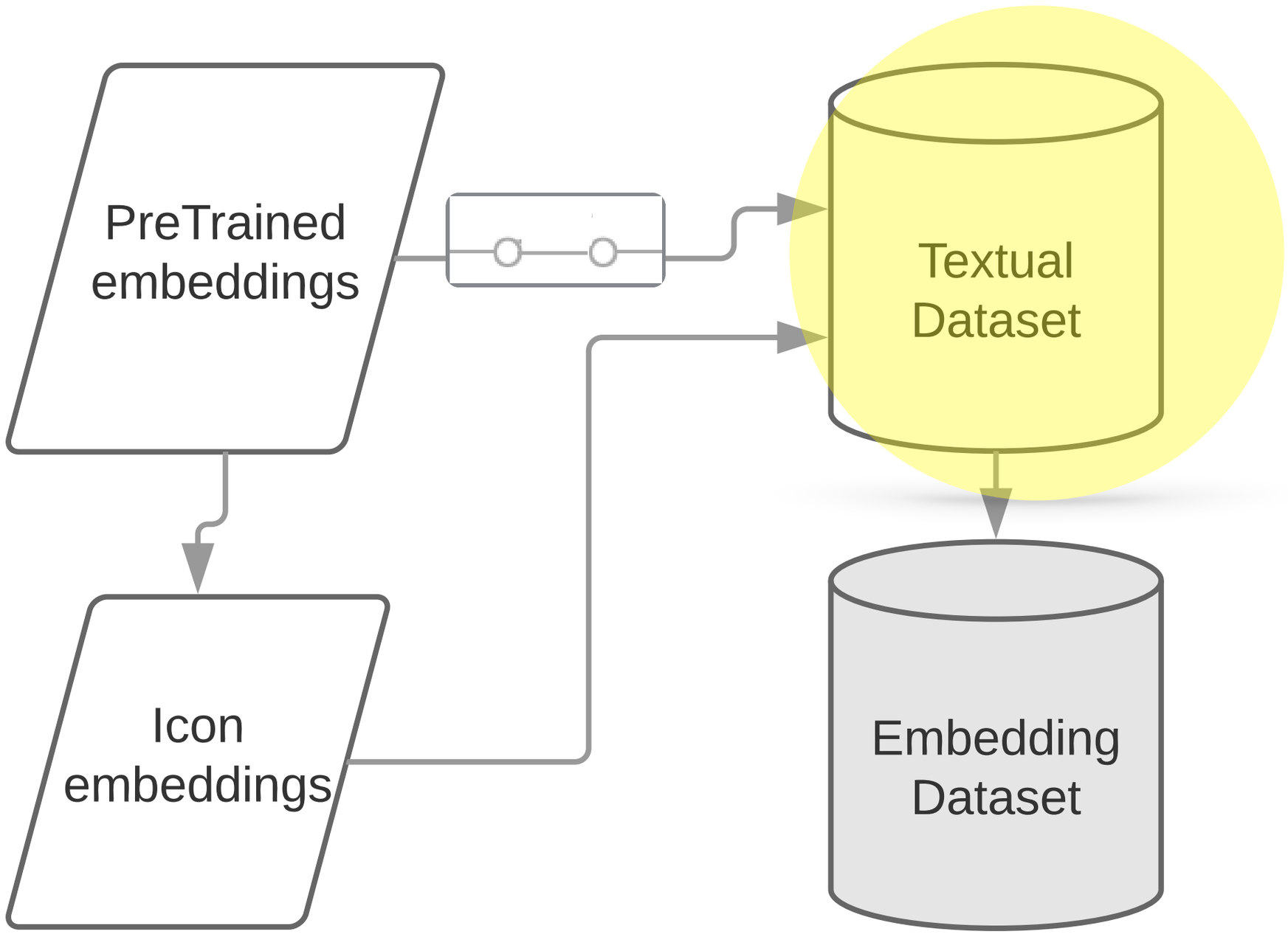
English: “so you did it”

non-pure:

<so> <you> <did> <it>

pure:

<so> <you> <it>


PreTrained
embeddings
Textual
Dataset
Icon
embeddings
Embedding
Dataset

Experiment 1+2:
Pretrained, Icon constraint

Glove	non-pure
MRR	0.85
ACC@1	49.29
ACC@10	92.29

c2v	non-pure
MRR	0.85
ACC@1	50.99
ACC@10	90.51

Experiment 1+2: Pretrained, Icon constraint

Glove	non-pure	pure
MRR	0.85	0.33
ACC@1	49.29	45.72
ACC@10	90.29	54.29

c2v	non-pure	pure
MRR	0.85	0.33
ACC@1	50.99	46.79
ACC@10	90.51	54.92