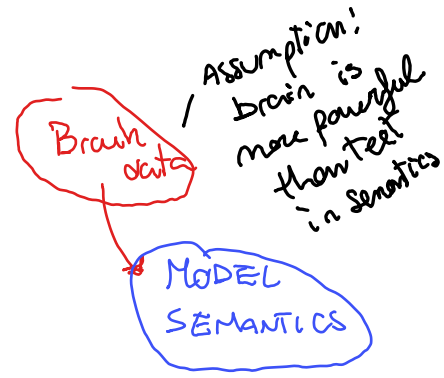
INTERPRETABLE SEMANTIC VECTORS FROM A JOINT MODEL OF BRAIN-AND TEXT-BASED MEANING

Fisher et al

- **VSM:** vector space model → represent semantical meaning by assigning each word a point in high dimensional space. Ex vectors from corpus. Distance between semantics maps into vectors' distance

  eg: count # occurences of a particular word near another

IDEA: use it with brain activations to improve semantic models



DATASET: → human-annotated word semantics
→ brain-informed VSM (fMRI & MEG for 18 subjects → read 60 nouns across 12 categories)

PREDICTION → infuse brain data to increase performance

MAP → map semantic concepts into the brain by jointly learning neural representation

CORPUS → 16 billions words corpus data

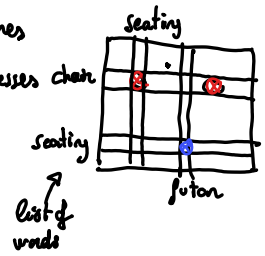
METHODS:

$X = \begin{matrix} w \\ w \end{matrix}\begin{bmatrix} 0 & \cdots \\ \vdots & \end{bmatrix}$ — co-occurence of w_1 with w_2 → symmetric → SVD → $w\begin{bmatrix} f \end{bmatrix}$

NNSE (NON NEGATIVE SPARSE EMBEDDING) → $\underset{A,D}{\operatorname{argmin}} \sum_{i=1}^{w} \| X_i - A_{i,:} \times D \|^2 + \lambda \| A \|_1$ (words; $A_{i,:} \geq 0$, $\|D\| \leq 1$)

↳ FIND BEST SCORING DIMENSION FOR EACH WORD (ROW)

eg: dim 1 for chair: seating, couches
dim 2: futon, mattresses chair



$A = \begin{matrix} w \\ \text{words} \end{matrix}\begin{bmatrix} \text{score} \\ \\ \\ \end{bmatrix}$ latent dimension

↳ cattura quando è importante una latent feature

JNNSE → Adding brain activity

$$\underset{A, D^{(a)}, D^{(b)}}{\operatorname{argmin}} \sum_{i=1}^{w} \| X_i - A_{i,:} \times D^{(a)} \|^2 + \sum_{i=1}^{w} \| Y_i - A_{i,:} \times D^{(b)} \|^2 + \lambda \| A \|_1$$

($A_{i,:} \geq 0$, $\|D\| \leq 1$)

per avere anche pochi dati } PARTIAL PAIRED DATA

Y_i: activity per word (just 60 words) $\times$ voxel

mappa latent dimension nelle parole $e \times f$

Pros: • handling partial paired data
• merge different brain imaging

- give constraints in A and D
- get best solution possible → we can specify

$A = \begin{matrix} w \\ \text{words} \end{matrix}\begin{bmatrix} \\ \\ \\ \end{bmatrix}$ loading of latent dimension z → they want predefined # of clusters

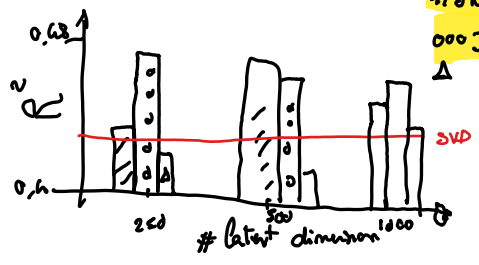
$D = z\begin{bmatrix} \text{score} \\ \\ \\ \end{bmatrix}$ c_o

Steps:

- matrix value must to be sparse (for interpretability)
- we can exploit $D^{(b)}$ and $D^{(c)}$ to predict

1. try to predict behavioral measure using VSM_{sa}
2. try to predict the 60×218 matrix from $60 \times A$ using JNNSE, NNSE

semantic annotation



- JNNSE (fMRI + Text)
- JNNSE (MEG + Text)
- NNSE (Text)

} increasing result

SVD → ~10% more similarity using brain data

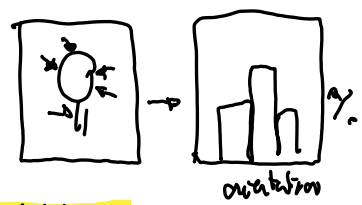
USING HUMAN BRAIN ACTIVITY TO GUIDE ML

Fang et al.

GOAL: improve representational geometry of ML methods

MODELS:

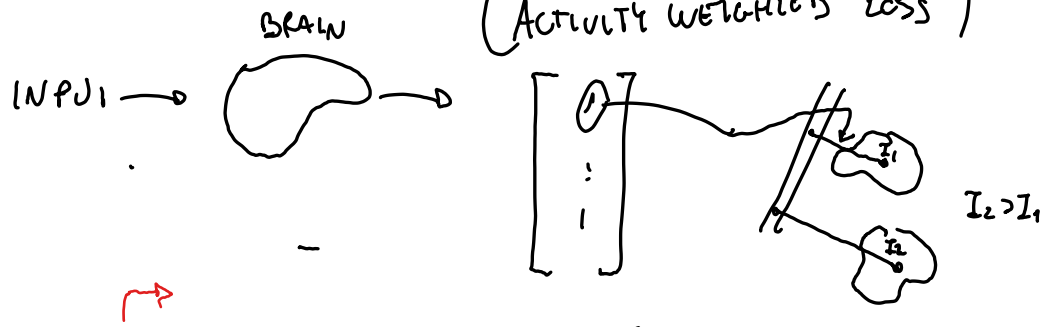
HOG
(Histogram of oriented gradients)

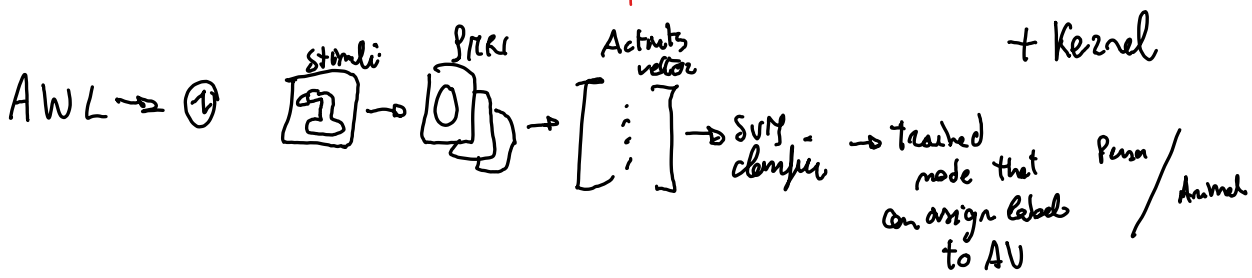


They consider only visual-related areas

CNN

DATA PRE PROCESSING:

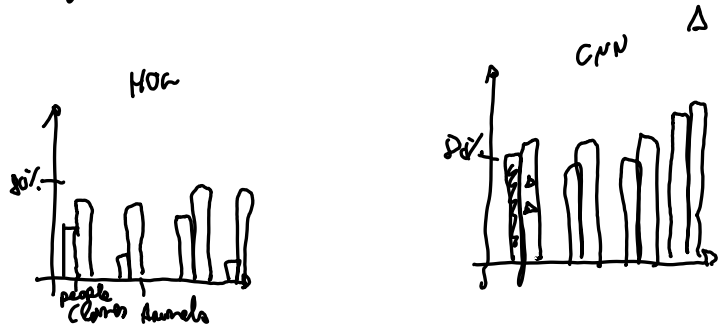




ADDING OF:
② • Radial basis function
• Margin renormalizing

e.g. Hinge loss
Δ AWL

RESULTS



CONCLUSIONS:

ML can be improved by brain data

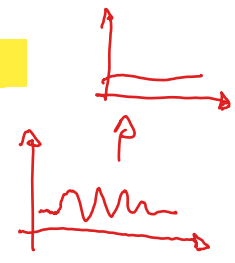
SURPRISE

LEARNING AND PREDICTION (Hamon)

SURPRISE-RESPONSE: depends on knowledge + low-level perception + accumulation of evidences biases noise

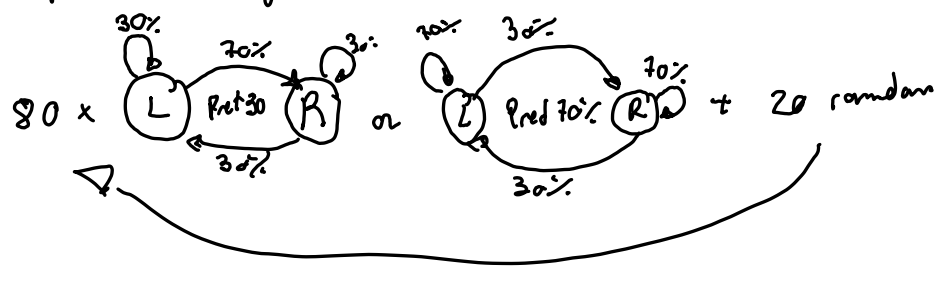
↳ STUDY OF EXPECTATION: probe the state of a cognitive system prior to stimulus presentation and independent of stimuli-guided responses

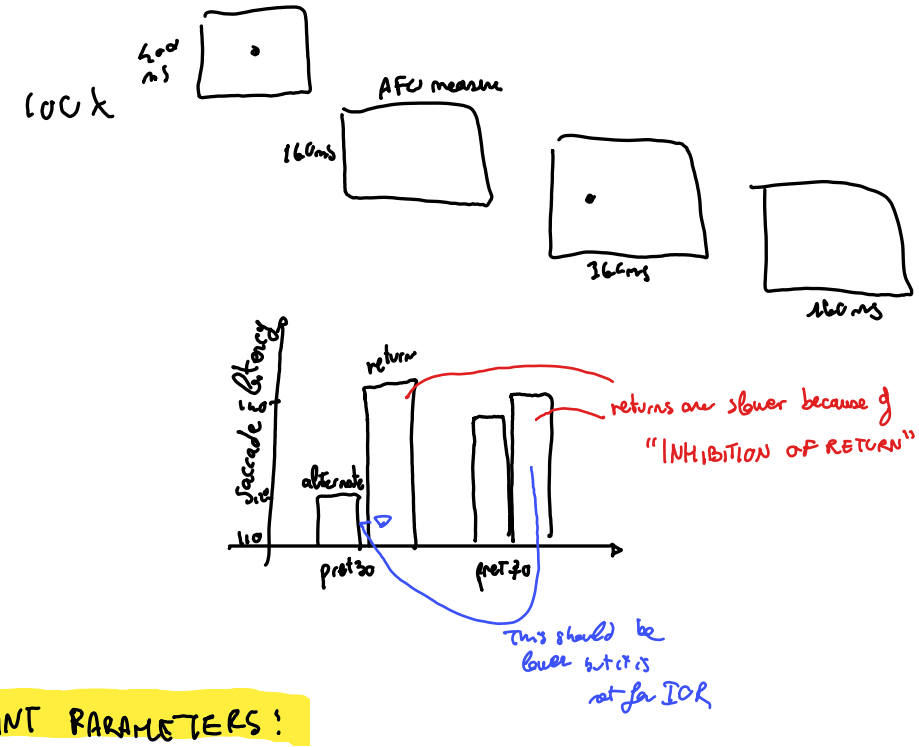
Previous works showed that presenting predictable stimuli have an impact on alpha waves (after some time the brain signal decrease → LEARNING) MEG



DATASET: N=21 volunteers

2 series → 200 samples with target position depending on Markov process (Pret 70, Pret 30)



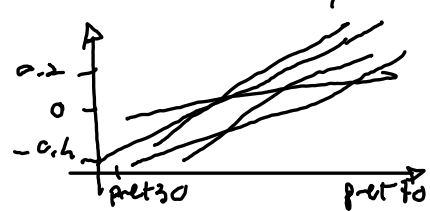


Two IMPORTANT PARAMETERS:

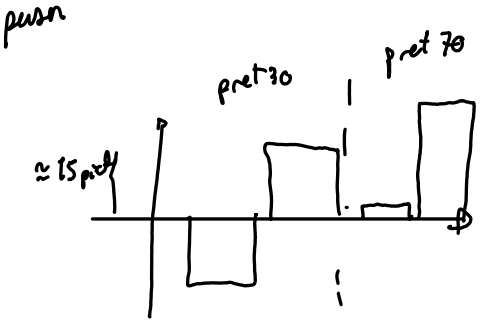
1. AGO (Anticipatory Gaze offset)
2. AFO (Anticipatory Fixation offset) → recording of gaze offset in relation to prior target measured last 10ms of pretarget blank screen

↓

We can plot "fixation density"

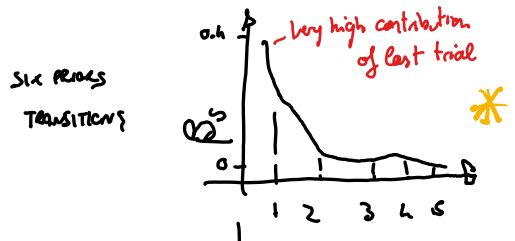


1 line per person



VALIDATION → INTERNAL (SPLIT-HALF) RELIABILITY: deriving two separate ΔAFO (indicator of learning) per participant: one for ODD and one for EVEN trials $SH = C, 9$

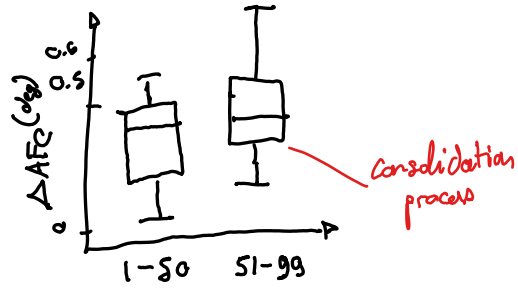
TIME SCALE OF LEARNING: we tend to do not predict the most likely outcome → IRRATIONAL: we should always bet on the most likely
↓ bias on last results



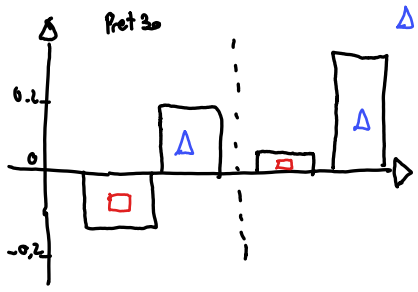
$\beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 \ldots + c + \varepsilon$
S=0 alternation
S=1 return

↓
This confirms statistical learning (high importance of last 3-5 transitions)

TIME SCALES OF KNOWLEDGE CONSOLIDATION AND FORGETTING: verified looking at bets on the first split (S0) and the second (S0)



WASHOUT: Look at bets on the last 20 random trials



→ Patients continues using their strategy to predict during the washout serie.

↓

HOW CAN WE MODEL THIS BEHAVIOUR?

o RE SCORLA-WAGNER learning model

"error driven learning model" → we can modify α according to *

learning rate

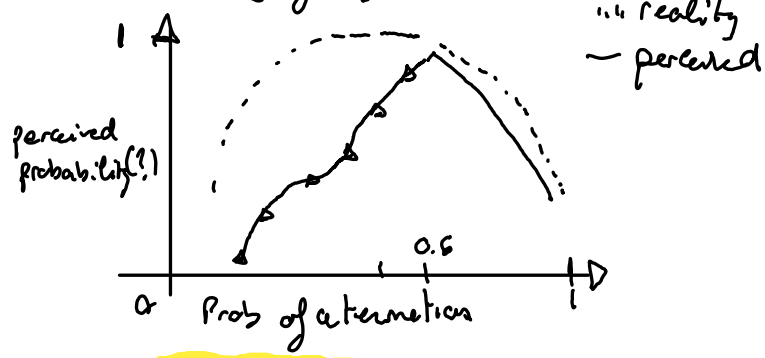
$$\begin{cases} P_{ret}(t+1) = P_{ret}(t) + \alpha(1 - P_{ret}(t)) & \text{after a return} \\ P_{ret}(t+1) = P_{ret}(t) - \alpha P_{ret}(t) & \text{after alternation} \\ AFC(t+1) = K \cdot P_{ret}(t+1) & \end{cases}$$

K: scaling factor transforming internal probability to overt behavior

eg: $P_{ret}(t) = 0.6$, $\alpha = 0.1$
ret → $P_{ret}(t+1) = 0.64$
alt → $P_{ret}(t+1) = 0.54$

Problem: for ret > 0.5, people will stout predicting returns because prob. exceed chance while in reality we mis-estimate chance → equilibrium at $P_{ret} = 0.6$

studying equilibrium



We can now correct the Rescorla Wagner model with this information

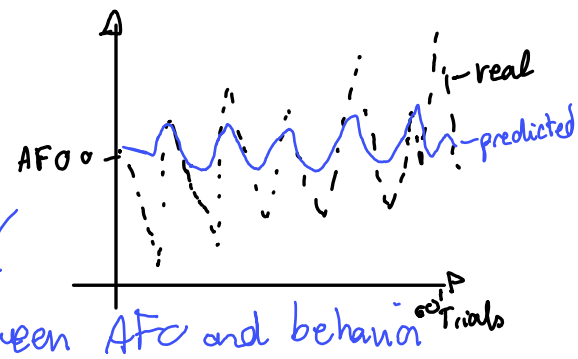
$$\begin{cases} P_{ret}(t+1) = P_{ret}(t) + \alpha\left(1 - P_{ret}(t)\right) & \text{after return} \\ P_{ret}(t+1) = P_{ret}(t) - \alpha P_{ret}(t) & \text{after alternation} \\ AFO(t+1) = K \cdot \left(P_{ret}(t+1) - P_0\right) & \end{cases}$$

P_0 → probability equilibrium point (bias)

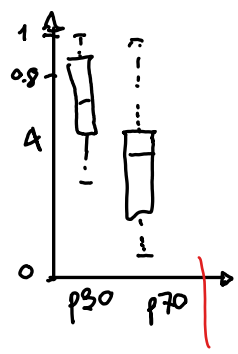
Now that we have a good model ⇒ TRAIN

↓

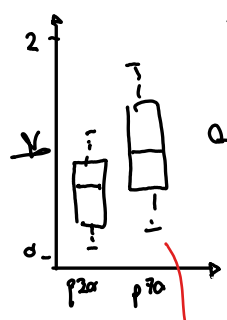
Leave one out (19/20 participants)

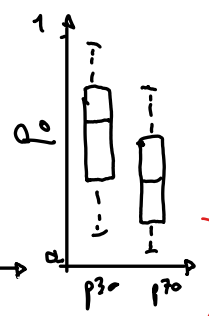


Strong relation between AFO and behavior



larger integration windows





stronger behavioral impacts while less bias (no stochasticity for humans)

↳ Questions:

→ If AFO measure prediction:

Larger AFO (>0) → faster return saccade and slower alternation saccade

↓ verified by the experiment

→ Which measure captures more info about predictions? AFO or SL (saccade latency)?

We can answer considering MUTUAL INFORMATION (MI) → consider the variation of uncertainty knowing another variable

$$I(x,w) = \sum_{x \in X} \sum_{w \in W} p(x,w) \log\left(\frac{P(x,w)}{P(x)P(w)}\right)$$

Result: AFO conveyed about twice as much information about the statistical process wrt SL

CONCLUSIONS: → ◦ Learning can be studied quantifying observable behavior over time

◦ macro-scale properties of the env. can be separated to micro-scale occurrences in recent past → Both impact behavior

◦ Learning translates into surprise signals → slow reaction to suprising events

◦ Learning produces anticipation/prediction and we can model it with error-driven learning

◦ Anticipatory signals can contain more inf. about the env. than stimulus related responses

SURPRISE AND NEGATION

Requirements: - Beliefs
- Attention to the world
- Appreciate contrast between event and beliefs
- suddenly, unexpectedly experience
- Novelty? → Def: not being previously experienced or encountered

NON NUOVO ≠ SORPRESA

Ekman & Davidson (1994) → surprise is an EMOTION that arises from mismatch between expectation and what is observed
- EMOTION: not beliefs but outcomes of beliefs about the world
- "VIOLATION OF EXPECTATION"

→ We can consider as a formal account of S as the "probability of the encountered event"
- as Frequency distribution

"Surprise is proportional to the expected probability of an encountered event

Problem: Surprise does not capture confidence → determined by # of evidence

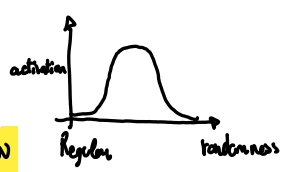
⇓

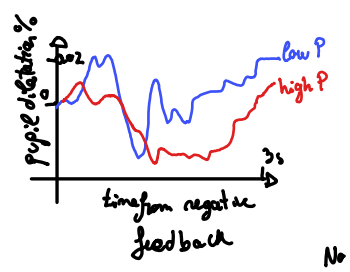
SURPRISE ← low probability
SURPRISE ← high confidence
SURPRISE ← emotional state (?)

→ internal motivator for interests and exploration
→ (Nordland 2003) → learning

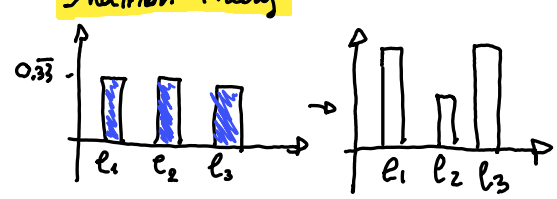
USE OF SURPRISE IN LEARNING → learning is proportional to surprise
- investigation of neural correlates of responses related to surprise

SURPRISE IN THE BRAIN



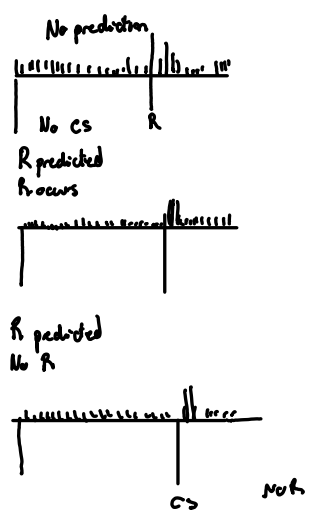


Shannon theory



Repetition priming: brain activity goes down when we have repetition

- Stimulus → Reward paradigms
(CS = stimulus learned to be followed by reward)



MULTIPLE SURPRISES AT DIFFERENT SCALES

→ Evidences that there is both { sensory specific surprise systems / general systems }

→ Some respond to surprise in a stimulus
→ other respond to overall surprise in input stream } these responses may track avg surprise linearly or non-linearly (inverse U)

EEG - • They can show increased activity for Cav - Shannon entrophy streams (many repetitions)

SURPRISE IN LANGUAGE (DISCUSSED)

↳ use EEG to measure neuroelectric activity at a very fine temporal scale
↳ words that are more "surprising" in context produce stronger response within a 5 sec

CONTRAST MODELS:

→ improbability and unexpectedness don't really mean the same
| likelyhood | depends on brain processes |

Es of roulette
→ 25/75% vs 25/27/38%
Es of balls extracted from a bag } whether or not you predict a different outcome seems to play an important role

$V_{Macedo}(k) = p_{max} - p(k)$ (most likely event)

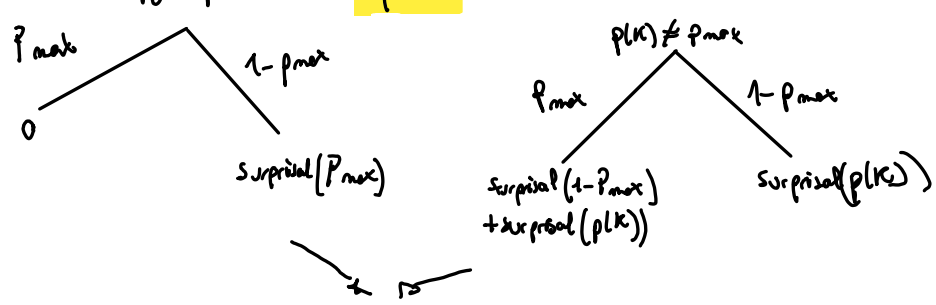
extremeness of p_{max}

$V_{bounded}(k) = 1 - \frac{\log_2(1-p(k))}{\log_2(1-p_{max})}$ → surprise is bigger when p_{max} is higher

$p(k) \neq 0.2$, $p(max) = 0.8$ → $1 - \frac{(-3)}{(-1)} = -2$

$p(k) = 0.2$, $p(max) = 0.4$ → $1 - \frac{(-3)}{(-2.3)} = \approx 0.8$??

Bounded surprise



a psychological model explains experienced surprise as a function of outcomes and p_{max}

SUMMARY:

- SURPRISE ≠ NOVELTY (expectation and deviation in context)
- if SURPRISE = PROBABILITY → ignore confidence and the role of alternative expectations
- CONTRAST MODELS → combine all together but are difficult to test
- Humans can code for surprise of a stimuli but also the avg surprise of an entire serie of stimuli (independent aspects of the environment)

e finalmente 

LANGUAGE COMPREHENSION AND INTEGRATION

→ we predict the future using learned association and recent past
→ Question: do we store language associations? And how do we use them for comprehension?

① From experience we memorize structures (words or type of phrases)
② we learn abstract rules that specify constraints on grammaticality
③ we acquire statistical knowledge on co-occurence of language entities

Behavioral psychology: S-R models for language → word is stimulus
Language comprension: the learned sequences of codjacent elements are internally represented as automatically characterizing sentence as it is encountered } NO EXPLANATION OF MEANING

BEHAVIORISM: Knowledge of language is knowing which words follow another word

THE S-R APPROACH (which word comes after another) House → traces → resolve just ① and ②

1 The → X → Y-es (art, n, v) ①'
2 The → Xes → Y- (det, n, v)

Acceptance of some level of abstraction → discourse structures → sentence → longer phrases → combination of shorter ones → single words

eg: "NP_1 V+ ed NP_2 ↔ NP_2 was V+ ed by NP_1" ②
- kernel declarative constructions
- passive sentence
- transformed into sentence
- Trasformation → what we learn

TWO PARALLEL IDEAS

+ other views SIGN, ASSIGN
zebra = horse + stripes

SUBSTITUTION VIEW
Signs (words) substituted for the objects in the world and both produce or similar reaction

DISPOSITION VIEW
Language produces a tendency (disposition) towards certain behaviors

CHOMSKY'S CHALLENGE TO S-R LEARNING (ABSTRACT RULES) 1959

Multiple weakness:

1. inability to explain comprehension of sentences not heard previously
2. inability to handle hierarchical structure and long-distance dependencies

↓

TRASFORMATIONAL GRAMMAR MODEL

→ GRAMMAR-RULE DRIVEN COMPOSITIONALITY → acoustic representation improves when words are not independent

↓ what is the keypoint?

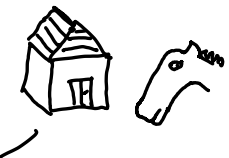
SENTENCE MEANING IS CAPTURED BY ITS STRUCTURE

↓

DERIVATIONAL THEORY OF COMPLEXITY → abandoned because of criticisms of method

PARSING BASED ACCOUNT → trying to assign a structure to what we have read/heard

↳ try to explain why some interpretation are preferred



- ⓐ Fails because of an human bias: we preferably associate words with the previous clause
- ⓑ Fail because of trace's memory demand

this refreshes the pointed part

LEVY

THEORY OF SYNTACTIC REVISION AND SURPRISAL

→ incremental parsing is done including constrains on potential continuations

→ parsing is resource based and act in parallel

→ includes the role of expectation which consists on ranking possible continuations based on experience

- ⓐ distributions change ($W(1 \dots w-1)$ vs $W(1 \dots w)$) can be studied with Kullback Liebler divergence
- ⓑ levy found out that surprise $\simeq -\log(P(\text{word} \mid \text{context}))$

→ we don't need to predict but just a MODEL for each word in time

PCFG

→ incoming words sharpen next-constituent expectations and thereby decreases surprisal at the final verb itself

A phrase is difficult is related just to how expensive is to integrate it into prior context

FRANK AND BOD'S

→ Difficulty = prob. without syntax

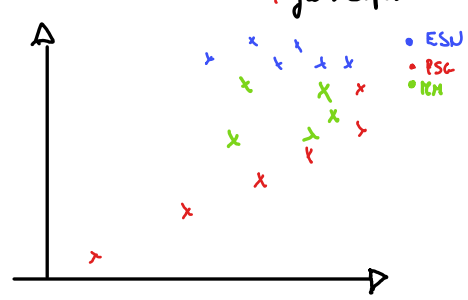
3 Test → PSG (phrase structure grammar) → They tried 4 combinations

- Markov Process (N-gram)
- Echo-state Network

Test how surprising is the truth → surprise → bad accuracy

Result → based on → PSYCHOLOGICAL ACCURACY reading time

→ LINGUISTIC ACCURACY from corpus



RUSSO ET AL 2020

→ DATASET : Brain activity of listeners that listen to a story

→ METHOD : • Describe how surprising was each word
• Compute vector-space representation

$$\text{surprisal}(t) = -\log_{10} P\left(W_t \mid W_1, \ldots, W_{t-1}\right)$$ → set window to 10

meaning → adjacency of a word with 1 of 1000 anchors (like Mitchell)

$$V_i = p(c_i|w)/p(c_i)$$

$$\text{Semantic weighted surprise} = p\left(W_T \mid W_{T-1}, \ldots, W_{T-10}\right) \sum_i w_i h_i \, p(c_i)$$

(area value; semantic vector for t-3, t-6?; scaling / Beta)

from corpus — 2ª order Markov-based surprise

semantic weight using anchors

↓

Regression Model

→ Predict brain activation given word and context

↓

Compute semantic weighted surprise

outcome → good prediction

Test 1 → remove probability → result is bad
Keeping only semantic part

Test 2 → remove conditional prob. → acceptable result
Keeping a raw probability

→ To conclude: might be difficult to dissociate different surprise models as they may be highly correlated

SwS and LS were correlated at $r = \pm 0.77$