Comparison of neural activity for appreciation of Japanese tanka in human brain and artificial intelligence

Shotaro Shiba Funai　@ Araya Inc. / QUP

Collaborators: Satoshi Iso (KEK / QUP), Junichi Chikazoe (Araya Inc.), Daichi Mochihashi (ISM, 統数研), Masayuki Asahara (NINJAL, 国語研), Naokazu Goda (NIPS, 生理研), Teppei Matsui (Okayama Univ. / JST PRESTO), Yutaka Shikano (Gunma Univ. / Chapman Univ. / JST PRESTO), Hirono Kawashima (Keio Univ.)

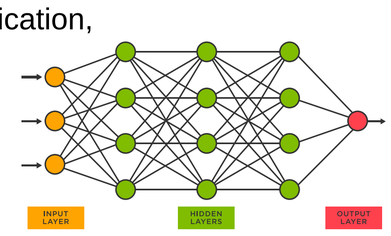
Our motivation: what is beauty?

- Beauty in physics: one of the projects in QUP
- Beauty is subjective experience but can be discussed in an objective way these days: foundation of neuroaesthetics, Prof. Zeki's talk (tomorrow).
- This progress is based on recent development of fMRI (brain activity) measurement and machine learning (as a method of artificial intelligence).

Machine learning

- Machine learning has an artificial neural network with layers, whose structure is similar to human brain.
- Linguistic machine learning is now applied for text classification, summarization, translation among various languages, …
- It seems to understand not only meaning of words but also context of sentences.

e.g., BERT *[Devlin et al. 2018]*, GPT *[Radford et al. 2018]*



https://www.tibco.com/

Our main question

Does the internal state of linguistic machine learning correspond to human brain activity when machines and humans read verse sentence with indirect implications?

(with beauty, sometimes)　(nontrivial context)

→ We chose Japanese tanka as an example of verse sentence.

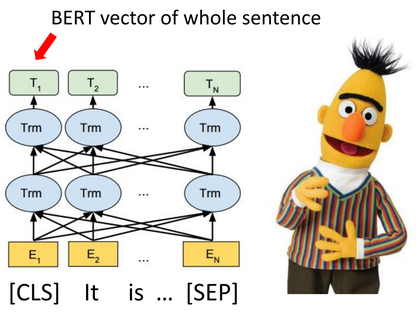
- It is a short poem with only 31 syllables.
- It often has meanings beyond its literal sense to make us emotional.

「研修中」だったあなたが「店員」になり真剣な眼差しがいい

You were "in training" and become a "clerk" then have good intent look.

BERT vector representations

- We use BERT (a popular linguistic machine learning) with the pretrained model (cl-tohoku/bert-Japanese, whole-word-masking, bpe).
- This model is trained with Japanese articles in Wikipedia: most of the sentences the machine learned are non-poetic.
- BERT has a word embedding layer + 12 encoder layers, and outputs 768-dimensional vector representations.
- The first vector (for [CLS]) is usually regarded as the representation of the whole sentence.
- Some researchers claim that shallow layers grasp syntactic properties while deep layers grasp semantic ones (but no consensus yet). *[Tenney-Das-Pavlick, 2019]*



https://medium.com/@cdathuraliya/

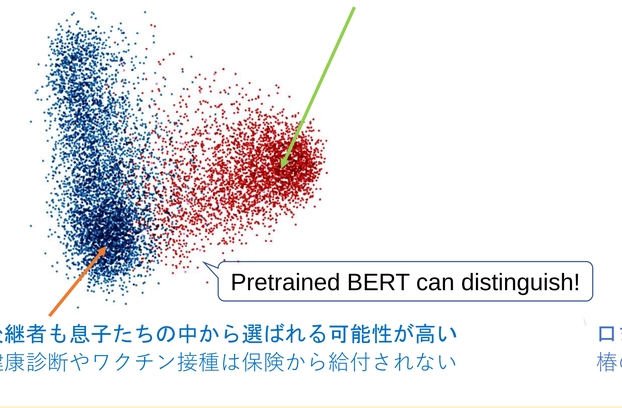
Using the BERT vectors, we picked out tankas and nonverse sentences which are relatively similar but can be mostly distinguished.

Our database (from NINJAL) → picked out → 150 tankas + 150 nonverse sentences

used for our experiment

美しき声より孤児となりゆきし少女なるべしいま秋立つ
あふれつつ四国の海の鳴る夜を汝が追憶は断たねばならぬ

好きだったあなたのくれたハンカチの匂い月面着陸をした
すでに豊作を報じる者とかかわりなく一本一本稲植える農婦



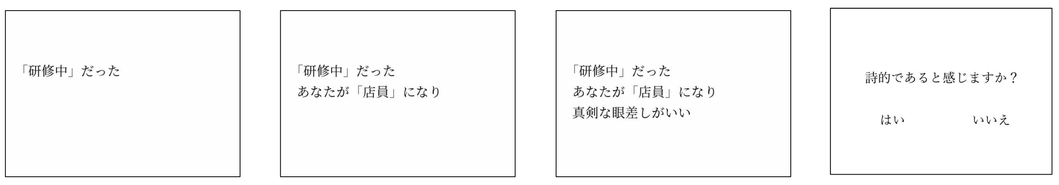
後継者も息子たちの中から選ばれる可能性が高い
健康診断やワクチン接種は保険から給付されない

口まわりの感覚が実に繊細になるのがひしひしとわかる
椿の花そして島の中心にある活火山の三原山

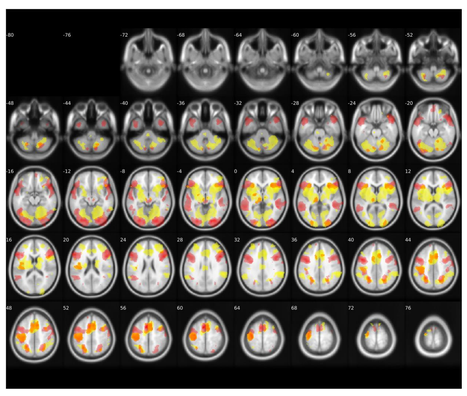
fMRI experiment

- Participants: 32 healthy young adults
- Functional Imaging: 3.0T scanner (at NIPS)
  GE-EPI, TR = 0.75s, TE = 31ms, flip angle = 55°,
  72 slices, multiband factor = 8, voxel size = 2.0 x 2.0 x 2.0 mm
- Preprocess: Realigned, slice timing corrected, normalized, using SPM12

We divide a tanka / sentence into 3 lines and show them with interval of 3 seconds.
After that, we ask the participants if they feel it is poetic or not ("judgement").



Result of fMRI (brain) measurement



vmPFC ~ mOFC, which plays a central role in neuroaesthetics.

Red: reaction to tanka stimulation

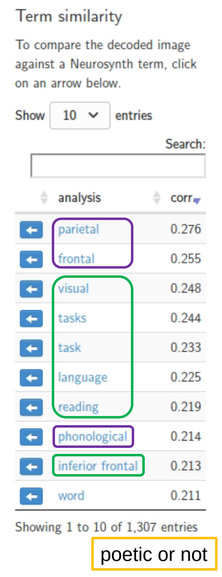
Language area and visual cortex (as expected)

Yellow: difference of reaction between poetic and non-poetic

Neurosynth suggests terms related to language processing and cognition, as expected.

Notable brain activations were found in precuneus, ventromedial PFC, left temporoparietal junction.

This may show that interaction of emotion and cognition takes place in these regions.



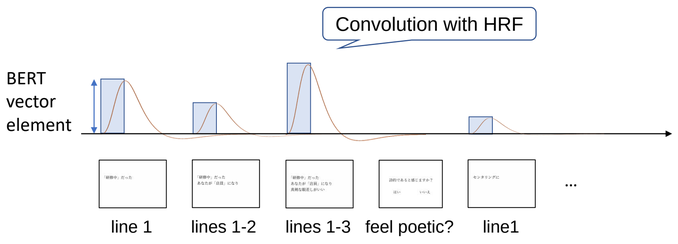
Correspondence of fMRI results and BERT layers

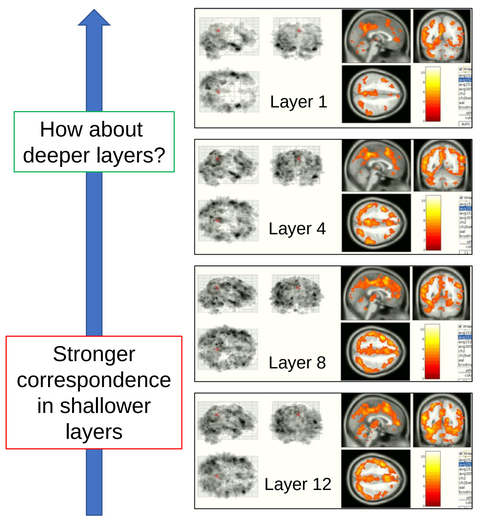
BERT vectors (768 dims)　trained with "judgements"

- For line 1, lines 1-2, all 3 lines of a tanka/sentence
- In layer 1 (deepest), 2, 3, …, 12 (shallowest)

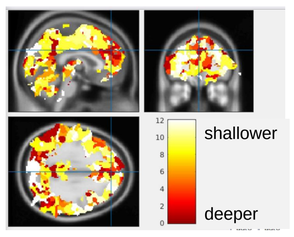
↓ PCA (reduction to 100 dims) and use elements in each direction

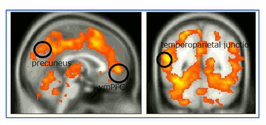
Regressors for regression analysis of fMRI data





We can expect that deeper layers of BERT correspond to brain area correlated with the judgement (poetic or not). Our result shows such correspondence!

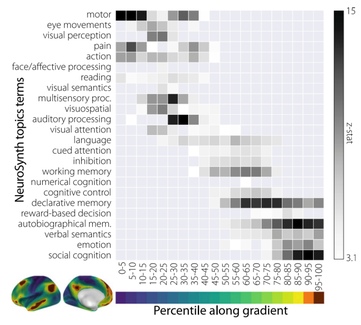


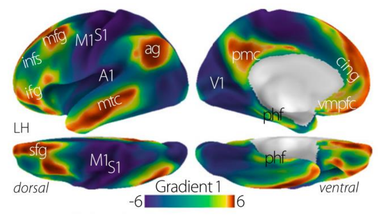


Brain area related to "poetic or not": Precuneus, ventromedial PFC, left temporoparietal junction, …

Correspondence with principal gradient (~brain's cognitive function)

concrete word ↕ abstract word



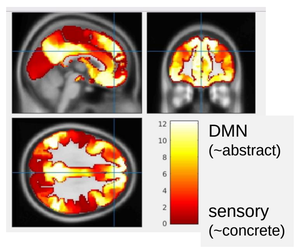


Blue regions process more concrete things.
Red regions process more abstract things.

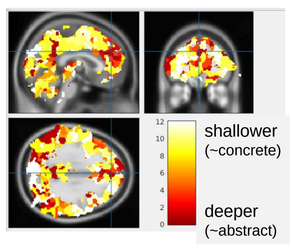
"Default mode network"　Margulies et al., 2016

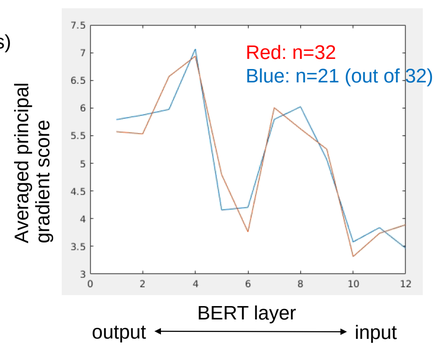
We can find reduce tendency of principal gradient from output (~abstract) to input (~concrete) in BERT. This is also a reasonable result!　But can we discuss what causes zigzags…?

Principal gradient (PG)



Our result (correspondence with BERT layers)





Summary

- ✓ We compared the neural activity in human brain (fMRI) and artificial intelligence (BERT) when the participants and the machines read Japanese tankas.
- ✓ We found that shallower layers of the pretrained BERT are strongly correlated with brain reactions in various area.
- ✓ We specified the brain area correlated with the judgements whether poetic or not. Then we find its correspondence to deeper layers of BERT (which presumably grasp semantic properties).
- ✓ We also show the correspondence of BERT layers and the principal gradient, which clarifies the hierarchical structure of brain from concrete to abstract cognitive functions.