Combining text mining and coordinate-based meta-analysis

Finn Årup Nielsen

DTU Compute
Technical University of Denmark

Workshop: Neuroimaging meta-analysis methods

April 17, 2015
Brain atlas constructed from abstract
Brain atlas constructed from experiment labels

1: Hot pain, Thermal pain, Warm temperature sensation
2: Finger movement, Localized movement, Motion, movement, locomotion
3: Externally generated threat response, Externally generated emotion, Threat
4: Memory, Cognition, Memory retrieval
5: Emotion, Mental process, Unpleasantness
6: Language, Rhyme judgement, Phonetic processing
7: Somesthesia, Perception, Cold pain
8: Face recognition, Objects (processing), Visual object recognition
9: Audiovisual speech perception, Multimodal perception, Congruent multimodal perception
10: Vision (visual perception), Reading, Saccadic eye movements
11: Self-reflection, Self processing, Self/other processing
12: Voice perception, Audition, Spatial neglect
13: Verbal fluency, Productive language, Silent word generation
14: Fear, Anger, S allele of promoter region of serotonin transporter gene
15: Syllable counting, Receptive language, Novelty seeking
16: Awake resting with eyes closed, Relaxed conscious state, Conscious state
17: Visuospatial attention, Visuospatial expectancy, Visuospatial processing
How is this done?
## Coordinate-based database

<table>
<thead>
<tr>
<th>Name</th>
<th>‘Old’ BrainMap</th>
<th>Brede</th>
<th>Neurosynth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry</td>
<td>Manual</td>
<td>Manual</td>
<td>Automatic</td>
</tr>
<tr>
<td>Collection</td>
<td>Web scraping</td>
<td>XML</td>
<td>TSV</td>
</tr>
<tr>
<td>Papers</td>
<td>224</td>
<td>186</td>
<td>≈ 10,000</td>
</tr>
<tr>
<td>Experiments</td>
<td>771</td>
<td>586</td>
<td>&gt; 18,000</td>
</tr>
<tr>
<td>Locations</td>
<td>7,263</td>
<td>3,912</td>
<td>&gt;150,000</td>
</tr>
<tr>
<td>Abstracts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Loc. labels</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Exp. labels</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>(Fox and Lancaster, 1994)</td>
<td>(Nielsen, 2003)</td>
<td>(Yarkoni et al., 2011)</td>
</tr>
</tbody>
</table>

Other databases: AMAT (Hamilton, 2009), SumsDB (Van Essen, 2009). Part of Brede Database is in AMAT. ‘New’ BrainMap.
Modeling of data in BrainMap

Gaussian mixture model for perception, cognition, motion (Nielsen and Hansen, 1999)

Kernel density estimation for audition, vision (Nielsen and Hansen, 2000)

\[ p(x|l): \text{x is Talairach space, } l \text{ is a BrainMap label ('behavioral domain')}. \]
Modeling anatomical labels in BrainMap

“We downloaded 7,263 location web-pages and 3,935 of these locations had an associated anatomical label” (Nielsen and Hansen, 2002b).

\[ p(x|l) \]: \( x \) is Talairach space, \( l \) is the anatomical label (word or phrase) associated with each coordinate, e.g., “occipital gyrus”, “gyrus”, “occipital”.

1,231 word/phrases.

Modeling anatomical labels in BrainMap

Kernel density estimation with leave-one-out cross-validation for determining the width of the kernel (Nielsen and Hansen, 2002b).

\[ E(\sigma^2, l) = - \sum_{n=1}^{N_l} \log p_{-n}(x_n|\sigma^2, l) \]

\[ p_{-n}(x|\sigma^2, l) = \frac{1}{N_l - 1} \sum_{n' \neq n}^{N_l} (2\pi\sigma^2)^{-3/2} \exp \left( -\frac{1}{2\sigma^2}(x - x_{n'})^2 \right) \]

Lars Kai Hansen’s algorithm with Newton optimization for optimizing the kernel width. Available in Brede Toolbox (Matlab).
Modeling anatomical labels in BrainMap

Finding outliers in BrainMap by looking at coordinates located in areas with low probability density.

Here $p(x|l = \text{cerebellum})$ with a coordinate in the anterior part of the brain

Typos: centimeter/millimeter, left/right, ... 

Algorithm also run for the Brede Database.
Brede Database

Matlab graphical user interface program for data entry and visualization of studies with Talairach coordinates.
Text mining and CBMA

Data entry of brain coordinates

Neuroimaging article

XML storage

3D visualization

Kernel density

NMF component

Matrix of kernel densities

Non-negative matrix factorization

One row
Brede Database neuroanatomy taxonomy

Hierarchy of brain regions.

Based on another neuroanatomical database ‘BrainInfo/NeuroNames’ (Bowden and Martin, 1995) and atlases, e.g. ‘Mai atlas’ (Mai et al., 1997).

Fields recorded: Canonical name, variation in names, abbreviations, links to NeuroNames and other databases.

Graph constructed with GraphViz (Gansner and North, 2000).
Text mining and CBMA

Functional segregation in brain regions

For a brain region = 1 to 313 brain regions:

Step 1: Get all coordinates for the specific area, build a density model, exclude coordinates that are outliers

Step 2: Determine themes of the brain area with text mining on abstracts that contain coordinates within the brain area

Step 3: Determine whether specific themes are spatially clustered in the brain area by testing whether two sets of coordinates are separated.

end

Step 4: Intertwine results from all brain regions

(Nielsen et al., 2006; Nielsen et al., 2005).
Step 1: Identify coordinates

Simple SQL-like command in Matlab to find locations

Corner cube visualization of 116 “posterior cingulate” coordinates found

An outlier: “Right postcentral gyrus/posterior cingulate gyrus” from (Jernigan et al., 1998).

Build kernel density estimate of the coordinates.
## Step 2: Bag-of words matrix

<table>
<thead>
<tr>
<th></th>
<th>‘memory’</th>
<th>‘visual’</th>
<th>‘motor’</th>
<th>‘time’</th>
<th>‘retrieval’</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fujii</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>...</td>
</tr>
<tr>
<td>Maddock</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Tsukiura</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Belin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Ellerman</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

For the further analysis: Include all papers that contain one or more of coordinates found.

Representation of the abstracts of the papers in a bag-of-words matrix: (abstract × words)-matrix ≡ \( X(N \times P) \).

Exclude a large list of word: Anatomical, “stop words”, ...
Step 2: Non-negative matrix factorization

Non-negative matrix factorization (NMF) decomposes a non-negative data matrix $X(N \times P)$ (Lee and Seung, 1999)

$$X = WH + U,$$

where $W(N \times K)$ and $H(K \times P)$ are also non-negative matrices.

“Euclidean” cost function for

$$E^{\text{eucl}} = \|X - WH\|_F^2$$

Iterative algorithm (Lee and Seung, 2001)

$$H_{kp} \leftarrow H_{kp} \frac{(W^T X)_{kp}}{(W^T W H)_{kp}}$$

$$W_{nk} \leftarrow W_{nk} \frac{(X H^T)_{nk}}{(W H H^T)_{nk}}.$$
Step 2: ‘Medial temporal lobe’ NMF result
Step 3: Test spatial distribution

Extract locations from grouped papers.

Test if the spatial distribution of locations for a group is different from the distribution from another group with Hotelling’s $T^2$ or convex hull pelling permutation test.

All possible tests within a level of non-negative matrix factorization are performed.
Step 4: ‘Cingulate gyrus’

Coordinates associated with topics on: pain, painful vs. memory, retrieval
Combining text and coordinate directly

Construct a functional parcellation of the brain based on combined text and coordinate analysis (Nielsen et al., 2004; Nielsen, 2009).

Brede Database with neuroimaging papers.

Bag-of-words matrix from abstracts, scaling, and stop words elimination getting a \((\text{paper} \times \text{words})\)-matrix, \(X_1\)

Kernel density estimates from coordinates contained in the papers getting a \((\text{papers} \times \text{vertices})\)-matrix or \((\text{papers} \times \text{voxels})\), \(Y\).

Product matrix: \(Z_1 = Y'X_1\) getting a \((\text{vertices} \times \text{labels})\)-matrix.

Approximative non-negative matrix factorization: \(WH \approx Z_1\) getting a \((\text{vertices} \times \text{topics})\)-matrix, \(W\) and a \((\text{topic} \times \text{words})\)-matrix, \(H\).
Winner-take-all function on $W$ (the surface) and $H$ (the legend).
Brain function ontology

Experiment label (‘external components’, brain function ontology, cognitive ontology) organized in a directed graph (Nielsen, 2005).

This graph can be converted to an adjacency matrix.

An experiment (i.e. contrast) in Brede Database may be labeled with an item from the ontology.
Functional parcellation with cognitive ontology

Brede Database with neuroimaging paper.

Adjacency matrix of experiment label (‘external components’, brain function ontology) graph \( C \) propagated \( D = \sum_i (\lambda C)^i \).

Propagated experiment label (experiments \( \times \) labels)-matrix \( X_2 = TD \), where the (experiments \( \times \) experiment labels)-matrix \( T \) represents the (manual) label of each experiment.

Kernel density estimates from coordinates contained in the experiment getting a (experiments \( \times \) vertices)-matrix or (experiments \( \times \) voxel)-matrix \( Y \).

Product matrix: \( Z_2 = Y'X_2 \) getting a (vertices \( \times \) experiment labels)-matrix.

Approximative non-negative matrix factorization: \( WH \approx Z_2 \) getting a (vertices \( \times \) topics)-matrix \( W \) and a (topic \( \times \) words)-matrix \( H \).
Brain atlas constructed from experiment labels

Winner-take-all function on W (the surface) and H (the legend with items from the experiment labels/ontology).
Brede tools

Presently developing a Python Brede Package with features such as Features: Handling of data from Brede Wiki, Neurosynth, word lists, data sets, surfaces, . . .: https://github.com/fnielsen/brede

Brede Database: http://neuro.compute.dtu.dk/services/brededatabase/

Brede Toolbox (matlab): http://neuro.compute.dtu.dk/software/brede/

References


Text mining of PubMed abstracts for detection of topics in neuroimaging studies mentioning posterior cingulate. Subsequent analysis of the spatial distribution of the Talairach coordinates in the clustered papers.


