Knowledge discovery in neuroinformatics

Technical University of Denmark, DTU Informatics
Cognitive Systems Section
Neuroinformatics Research Group

“Coordinate-based meta-analytic search of neuroscientific literature and its expansion using semantic keyword extraction”

Speakers: BARTŁOMIEJ WILKOWSKI
MARCIN SZEWCZYK

Cimbi
Center for integrated molecular brain imaging

The Lundbeck Foundation

National Institutes of Health (NIH), 9000 Rockville Pike, Bethesda, Maryland 20892 – June 25, 2009
Neuroinformatics Research Group

Professor Lars Kai Hansen

Finn Årup Nielsen (Senior Researcher)

Bartłomiej Wilkowski (PhD Student)

Marcin Marek Szewczyk (Research Assistant)

Peter Mondrup Rasmussen (PhD Student)
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
Growing number of functional neuroimaging studies → demand for:

- Data integration,
- Data dissemination between research centers;

(Ascoli, 2006) – “The Ups and Downs of Neuroscience Shares”
(Teeters et al., 2008) - “Data Sharing for Computational Neuroscience”

Functional localization hypothesizes that a given human behavior is established by a change in brain activity in a relatively limited number of spatially segregated processing units → demand for:

- Efficient (coordinate/localization-based) searching of references to any related literature;
Project overview

- Develop the tools for meta-analysis and efficient searching of related literature/experiments given coordinate(s) in brain (knowledge discovery):
  - Database offering coordinate-based querying service
  - Software to facilitate literature searching directly from neuroscientists' common environments (SPM, FSL, ...)
  - Extending coordinate-based search results by querying bigger, more comprehensive databases like PubMed
  - Creating a secure web-service for neuroscience for stimulation of data and experience dissemination among research groups
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
Brede Database

- Close to 4000 coordinates from 186 papers with a total of 586 experiments
- Firstly, data stored in XML files. Recently, moved to MySQL database.
- Recording published neuroimaging experiments that list stereotaxic coordinates in so-called MNI or Talairach space (Talairach and Tournoux, 1988) - "Co-planar Stereotaxic Atlas of the Human Brain"
Coordinate-based searching in Brede DB

---

**Brede Database - Talairach coordinate search**

brede_loc_query — Search after locations (Talairach coordinates) in the Brede Database

<table>
<thead>
<tr>
<th>#</th>
<th>Distance</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>WOBIB</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.4</td>
<td>48</td>
<td>-36</td>
<td>8</td>
<td>130</td>
<td>Superior temporal gyrus — Tics during Tourette’s syndrome (WOEXP: 402)</td>
</tr>
<tr>
<td>2</td>
<td>6.1</td>
<td>47</td>
<td>-40</td>
<td>9</td>
<td>128</td>
<td>Right temporoparietal junction — Visuoproprioceptive conflict (WOEXP: 393)</td>
</tr>
<tr>
<td>3</td>
<td>6.6</td>
<td>48</td>
<td>-40</td>
<td>8</td>
<td>177</td>
<td>Middle and posterior temporal — Happiness from films and recall (WOEXP: 540)</td>
</tr>
<tr>
<td>4</td>
<td>7.1</td>
<td>49</td>
<td>-40</td>
<td>13</td>
<td>91</td>
<td>Right superior temporal — Alzheimer’s disease versus healthy (WOEXP: 291)</td>
</tr>
<tr>
<td>5</td>
<td>7.2</td>
<td>45</td>
<td>-31</td>
<td>17</td>
<td>39</td>
<td>Unpleasant words (WOEXP: 132)</td>
</tr>
<tr>
<td>6</td>
<td>8.5</td>
<td>43</td>
<td>-26</td>
<td>10</td>
<td>64</td>
<td>Right Heschl’s gyrus — Listening to voices (WOEXP: 199)</td>
</tr>
<tr>
<td>7</td>
<td>8.6</td>
<td>52</td>
<td>-38</td>
<td>7</td>
<td>168</td>
<td>Right superior temporal sulcus — Threat-related words in controls versus panic disorder patients (WOEXP: 515)</td>
</tr>
<tr>
<td>8</td>
<td>8.7</td>
<td>52</td>
<td>-37</td>
<td>7</td>
<td>86</td>
<td>Right middle temporal gyrus — Activation in sadness film viewing versus neutral film viewing (WOEXP: 282)</td>
</tr>
<tr>
<td>9</td>
<td>9.0</td>
<td>50</td>
<td>-30</td>
<td>16</td>
<td>59</td>
<td>Superior temporal gyrus — Spatial neglect (WOEXP: 185)</td>
</tr>
</tbody>
</table>
An fMRI experiment resulting in 29 reported coordinates

Brede Database offers:

- location search (distance between coordinates)

- 'experimental' search (similarity between two sets of coordinates / volumes)

(Nielsen and Hansen, 2004) - "Finding related functional neuroimaging volumes"
"Statistical Parametric Mapping refers to the construction and assessment of spatially extended statistical processes used to test hypotheses about functional imaging data. These ideas have been instantiated in software that is called SPM."

"The SPM software package has been designed for the analysis of brain imaging data sequences. The sequences can be a series of images from different cohorts, or time-series from the same subject. The current release is designed for the analysis of fMRI, PET, SPECT, EEG and MEG."

Taken from: http://www.fil.ion.ucl.ac.uk/spm/
BredeQuery plugin for SPM

http://neuroinf.imm.dtu.dk/BredeQuery/
The coordinates of the most significant activations in brain, found during an SPM analysis, are:

1. grabbed by the BredeQuery plugin,
2. transformed using any of MNI to Talairach transformations,
3. prepared for a coordinate-based searching with Brede Database;
MNI-to-Talairach transformations

- **brett** - Piece-wise affine transformation by Matthew Brett *(Brett, 1999)* - "The MNI brain and the Talairach atlas."

- **lancaster** – affine transformation by Jack Lancaster et al. *(Lancaster et al., 2007)* - "Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template."
  - SPM
  - FSL
  - POOLED (combined)
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
SKEEPMED

COORDINATES

Medical document → SKEEPMED → PubMed query

UMLS®

python powered

MetaMap

BREDE DATABASE

DTU

(brain_part_1 OR brain_part_2 OR ...) AND (term_1 AND term_2 AND ...)

RELATED PUBLICATIONS

PubMed
- Load text (abstract, article):
  - `skeepmed_input_xml = open(xml_file_path,'r')`

- Run MetaMap:
  - `metamap_file_exec_path = '/usr/local/bin/metamap08'`
  - `parameters = '-% format abstract.txt metamap_out_file.xml'`
  - `metamap_log = subprocess.Popen([metamap_file_exec_path, parameters],stdout=subprocess.PIPE).communicate()[0]`

- Parse MetaMap XML and `getListOfKeywords()`:
  - Check all Mappings and their Candidates, select those with sufficient NegScore, count frequency of each keyword occurrence, store in a dictionary (keyword:freq)

- Create query, ask PubMed
Keywords

- Two types of keywords:
  - brain_parts
  - terms
- Brain_parts retrieval settings:
  - Only Neuronames Brain Hierarchy data source used
  - Threshold low
- Terms retrieval settings:
  - All data sources used
  - Threshold high = 1000 (max) (only best matches)
  - Minimum occurrence frequency > 1
(brain_part_1 OR brain_part_2 OR ...) AND (term_1 AND term_2 AND ...)
Table 1: Results of spatial closeness comparison between experiments from PubMed retrieved articles and “source”. Column 3 shows the position of the best-matched “source’s” experiment in the results list returned by Brede Database when querying a test article experiment coordinates. In the parentheses the percentage of matched “source’s” experiments found in top 20 Brede Database results is shown.

<table>
<thead>
<tr>
<th>#</th>
<th>PubMed Article</th>
<th>Year</th>
<th>Position in Brede Database search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neural correlates of heart rate variability during emotion</td>
<td>2009</td>
<td>1 (70%)</td>
</tr>
<tr>
<td>2</td>
<td>Beyond disgust: impaired recognition of negative emotions prior to diagnosis</td>
<td>2007</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>3</td>
<td>Disgust and happiness recognition correlate with anteromedial insula and amygdala volume respectively in preclinical Huntington's disease</td>
<td>2007</td>
<td>3 (20%)</td>
</tr>
<tr>
<td>4</td>
<td>An event related functional magnetic resonance imaging study of facial emotion processing in Asperger syndrome</td>
<td>2007</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Neurophysiological correlates of induced discrete emotions in humans: an individually oriented analysis</td>
<td>2006</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>6</td>
<td>Neurophysiological correlates of induced discrete emotions in humans: an individual analysis</td>
<td>2004</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>7</td>
<td>Functional neuroanatomy of emotions: a meta-analysis</td>
<td>2003</td>
<td>9 (20%)</td>
</tr>
<tr>
<td>8</td>
<td>Common and distinct neural responses during direct and incidental processing of multiple facial emotions</td>
<td>2003</td>
<td>1 (10%)</td>
</tr>
<tr>
<td>9</td>
<td>A preferential increase in the extrastriate response to signals of danger</td>
<td>2003</td>
<td>1 (10%)</td>
</tr>
<tr>
<td>10</td>
<td>Impaired facial emotion recognition in early-onset right mesial temporal lobe epilepsy</td>
<td>2003</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>11</td>
<td>Age-related differences in brain activation during emotional face processing</td>
<td>2003</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>An fMRI study of facial emotion processing in patients with schizophrenia</td>
<td>2002</td>
<td>2 (60%)</td>
</tr>
<tr>
<td>13</td>
<td>Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI</td>
<td>2002</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>14</td>
<td>Deficits in recognition of emotional facial expression are still present in alcoholics after mid- to long-term abstinence</td>
<td>2001</td>
<td>no coordinates reported</td>
</tr>
<tr>
<td>15</td>
<td>Activation of anterior paralimbic structures during guilt-related script-driven imagery</td>
<td>2000</td>
<td>7 (50%)</td>
</tr>
<tr>
<td>16</td>
<td>Perception of emotion in frontotemporal dementia and Alzheimer disease</td>
<td>1999</td>
<td>no coordinates reported</td>
</tr>
</tbody>
</table>

Test coordinate: (-8,1,9) – thalamus brain region

Brede Database best match:
"Neuroanatomical Correlates of Happiness, Sadness, and Disgust" by Richard D. Lane et al. (1997)

Keywords:

**brain_part:** cerebral cortex, thalamus, insula, frontal lobe

**term:** disgust, sadness, happiness, emotion
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
Functionality evaluation

- How well works our recent pipeline?
- Need for automatic evaluation of the results – how? (current consultations with professor Ingemar Cox)
- Find the best Metamap parameters settings (data sources, semantic types, thresholds) – employment of metaheuristics?
- Combine data mining, machine learning, statistical methods (LSA, NMF, etc.) with ontological mapping?
Metaheuristics

- Thousands of parameters: threshold value (0..1000), 135 Semantic Types, 148 UMLS Sources → $2^{10} \cdot 2^{135} \cdot 2^{148} = 2^{293}$

- Metaheuristics used for finding the best parameters' setting (very stable results)

- Algorithm type: tuned simulated annealing

- 3 random articles for tuning, 3 random articles for testing

- Evaluation (golden set – 20 papers from PubMed)
Secure portal for neuroscientists
Secure portal for neuroscientists

- Integrated toolkit for encrypted communication
- Mixture of symmetric and asymmetric cryptography protocols to securely exchange information within virtual groups and public
- Version control
- Ability to securely exchange documents, coordinates
- Peer review system
- Ability to easily publish given work
Hopes for the future of MetaMap

- Unicode support
- Native 64-bit platform
- Ability to query for semantic types
- Ability to query for UMLS sources
Hopes for the future of MetaMap

- Both stand alone application and service oriented
- Ability to extract UMLS mapping hierarchy
  - parent, child
  - siblings, synonyms
- Open Python API
Roadmap

- Motivations and project overview
- Coordinate-based searching (BredeDatabase & BredeQuery plugin for SPM)
- Semantic KEyword Extraction Pipeline for MEdical Documents (SKEEPMED)
- Future directions, bottlenecks, problems
  - Validation and evaluation
  - Machine learning & ontologies (hybrid approach)
  - Metaheuristics for finding the best MetaMap parameters setting
- Conclusions
Thank you for your attention!

Questions?

Bartłomiej Wilkowski - bw@imm.dtu.dk