The Spatio-Temporal Representation of Natural Reading

Leila Wehbe
Machine Learning Department
Carnegie Mellon University
lwehbe@cs.cmu.edu

My thesis is about studying how the brain organizes complex information when it read text in a naturalistic setting. My work is an integrated interdisciplinary effort which employs *functional neuroimaging*, and revolves around the development of *machine learning* methods to uncover multilayer cognitive processes from brain activity recordings.

Studying how the human brain represents meaning is not only important for expanding our scientific knowledge of the brain and of intelligence. By mapping behavioral traits to differences in brain representations, we increase our understanding of neurological disorders that plague large populations, which may bring us closer to finding treatments (as detailed in the last section of this statement). For all these purposes, functional neuroimaging is an invaluable tool.

Traditional functional neuroimaging studies typically consist of highly controlled experiments which vary along a few conditions. The stimuli for these conditions are artificially designed, and therefore might result in conclusions that are not generalizable to how the brain works in real life. When studying language processing for example, very few experiments show subjects a real text, and show instead carefully designed stimuli.

Furthermore, the analysis of functional neuroimaging data has typically consisted in simple comparisons: regions which respond differently to the individual conditions are identified. Many researchers have recently started using *brain decoding* (i.e. classifying the stimulus being processed from the subject's brain image), which can reveal responses encoded in subtle patterns of activity across a brain region. However, brain decoding is still mostly used in a rather limited fashion. In order to predict which condition an image corresponds to, a classifier is trained on several examples of each condition. This classifier is *not* able to generalize its knowledge to *novel* conditions not seen in training. It can therefore be argued that such a model does not represent a broad understanding of brain function.

This work revolves around studying the parallel cognitive processes involved when subjects are engaged in a *naturalistic* language processing task, namely reading a chapter of a real book. We use computational linguistics algorithms to model the content of the text, and machine learning to identify regions in the brain that are involved in processing its different components. This work is consequently an integrated interdisciplinary effort.

The spatial representation of language subprocesses

We set out to challenge the understanding that it is difficult to study the complex processing of natural stories. We used functional Magnetic Resonance Imaging (fMRI) to record the brain activity of subjects while they read an unmodified chapter of a popular book. Unprecedently, we modeled the measured brain activity as a function of the content of the text being read Wehbe *et al.* [2014a]. Our model is able to extrapolate to predict brain activity for novel passages of text - beyond those on which it has been trained. Not only can our model be used for decoding what passage of text was being read from brain activity, but it can also report which type of information about the text (syntax, semantic properties, narrative events etc.) is encoded by the activity of every brain region. Using this model, we found that the different regions that are usually associated with language are processing different types of linguistic information. We are able to build detailed reading representations maps, in which each region is labeled by the type of information it processes in the distributed task of story understanding.

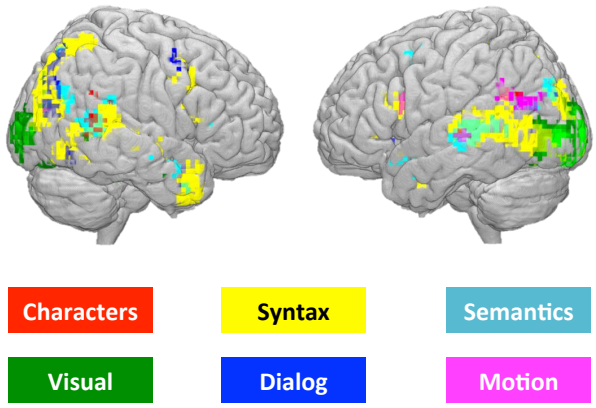

Figure 1: Brain map of the different reading sub-processes computed from combing data from multiple subjects. Each region is colored by the type of information it represents when subjects read a complex text. Details in Wehbe *et al.* [2014a].

Our approach is important in many ways. We are able not only to detect *where* language processing increases brain

activity, but also to reveal *what* type of information is encoded in each of the regions classically reported as responsive to language. From one experiment, we can produce a large number of findings. Had we chosen to follow the classical method, each of our results would have required its own experiment. Our approach makes neuroimaging much more flexible. Furthermore, if a researcher develops a new reading theory after running an experiment, they can annotate the stimulus text accordingly, and test the theory against the previously recorded data without having to collect new experimental data.

The time-line of meaning construction:

To study the sub-word dynamics of story reading, we turned to Magnetoencephalography (MEG), which can record brain activity at a time resolution of one millisecond. We collected MEG recordings when the subjects undergo the same naturalistic task of reading a complex chapter from a novel Wehbe *et al.* [2014b]. We were interested in identifying the different stages of continuous meaning construction when subjects read a text. We noticed the similarity between neural network language models which can "read" a text word by word and predict the next word in a sentence, and the human brain. Both the models and the brain have to maintain a representation of the previous context, they have to represent the features of the incoming word and integrate it with the previous context before moving on to the next word.

We used the language models to detect these different processes in brain data. Our novel results reveal that context is more decodable than the properties of the incoming word, hinting that more brain activity might be involved in representing context. Furthermore, the results include a suggested time-line of how the brain updates its context representation. They also demonstrate the incremental perception of every new word starting early in the visual cortex, moving next to the temporal lobes and finally to the frontal regions. Lastly, the results suggest that the integration process occurs in the temporal lobes after the new word has been perceived.

Methodology:

The cognitive science contributions are summarized above, however this thesis also consists of a series of projects in order to improve different parts of this complex pipeline. We undertook an extensive project that compared different methods for predicting fMRI data from feature annotations of the stimuli Wehbe *et al.*. Interestingly, it turns out that different regularization approaches result in comparable brain decoding performance. More importantly, when using single voxel (volume-pixel) data, there is a strong correspondence between classification accuracy and the tuning regularization parameter (picked by cross-validation). This finding has an important application in voxel-selection methods for brain decoding. Another example is a collaboration with other researchers to create interpretable vector space models of semantics composition, which outperformed other semantic composition models on multiple tasks Fyshe *et al.*.

We are currently working on shortcutting the decoding and hypothesis-testing approach. Most neuroimaging experiments fall either in the category of two-sample testing (establishing if a brain area responds differently to different stimuli) or in the category of independence testing (is the activity in an area independent of the representation of the stimulus in a feature space). Most approaches first estimate a statistic (e.g. the regression weight when a voxel's activity is fit to the stimulus, or the classification accuracy when using a decoding task), and then perform a hypothesis-test on this statistic. Instead of using these two steps, modern nonparametric hypothesis tests can be applied directly to these problems, and we have been working on adapting them to the high-dimensional and limited sample-sized brain data Ramdas* and Wehbe*.

References

Alona Fyshe, L Wehbe, P Talukdar, B Murphy, and T Mitchell. A compositional and interpretable semantic space. *Proceedings of the 2015 Conference of the North American Chapter of the ACL*.

Aaditya Ramdas* and Leila Wehbe*. Nonparametric independence testing for small sample sizes. *Proceedings of the 2015 International Joint Conference on Artificial Intelligence (IJCAI)*.

Leila Wehbe, Aaditya Ramdas, Rebecca Steorts, and Cosma Shalizi. Regularized brain reading with shrinkage and smoothing. *in review*.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9(11): e112575, 2014.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014. Association for Computational Linguistics.