Distinguishing Natural Language Processes on the Basis of fMRI-Measured Brain Activation

Francisco Pereira[1], Marcel Just[2], and Tom Mitchell[1]

[1] Computer Science Department
fpereira@cs.cmu.edu
[2] Psychology Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
just+@andrew.cmu.edu

Abstract. We present a method for distinguishing two subtly different mental states, on the basis of the underlying brain activation measured with fMRI. The method uses a classifier to learn to distinguish between brain activation in a set of selected voxels (volume elements) during the processing of two types of sentences, namely ambiguous versus unambiguous sentences. The classifier is then used to distinguish the two states in untrained instances. The method can be generalized to accomplish knowledge discovery in cases where the contrasting brain activation profiles are not known a priori.

1 A fMRI Study of Sentence Processing

1.1 Introduction

This paper builds on an fMRI (functional Magnetic Resonance Imaging [2]) study of cortical activity during the reading of syntactically ambiguous sentences [3]. The latter are sentences in which a word can have one of two lexical and syntactic roles, and there is no disambiguating information in the context that precedes the ambiguity. However, the ambiguity is resolved by information occurring later in the sentence. For instance, in

The horse raced past the barn escaped from his trainer.

the meaning of the sentence is clear after the word "escaped" is reached. The ambiguity occurs at "raced", which is could be interpreted as either a past tense (preferred) or a past participle (unpreferred). An unambiguous sentence could be, for example

The experienced soldiers spoke about the dangers before the midnight raid.

where "spoke" is unambiguously the past tense form.

L. De Raedt and A. Siebes (Eds.): PKDD 2001, LNAI 2168, pp. 374–385, 2001.

The study analyzed the activation in different parts of the brain every 1500 msec during the reading of ambiguous sentences and unambiguous control sentences. The analysis performed examined both the amount and location of such activation and the contrast between the two types of sentences, expressed in those terms.

One additional dimension of analysis could be the characterization of what is different between activation in the two experimental conditions. In addition to the amount of activation triggered, one might have to consider differences in the shape of the activation response, in localization (i.e. some points are only active in one of the conditions), and in timing. One could even consider the question of ascertaining whether there is more than one kind of cognitive process taking place.

But while it is relatively simple to test for such things as differing amounts of activation, that is not the case for the other questions. We propose a method for identifying specific locations in the brain where activation patterns are distinguishable across experimental conditions and, through that set of locations, allowing the discovery of answers to the questions above.

1.2 Syntactical Ambiguity Experiment

Let us take a closer look at what it means for a sentence to be ambiguous. The development of the sentence can be more or less surprising, and sentences taking the less likely meaning are called *ambiguous unpreferred* sentences. Ambiguous sentences which develop with the most predictable meaning are called *ambiguous preferred* sentences, and sentences without any ambiguity are *unambiguous* sentences. We shall concentrate on distinguishing ambiguous unpreferred and unambiguous sentences, and hence shall use the designations ambiguous/unambiguous from this point onwards.

The study above concentrated on two cortical areas known to be involved in sentence processing, the Left Inferior Frontal Gyrus (LIFG), also known as Broca's area, and the Left Superior Temporal Gyrus (LSTG)/ Left Posterior Middle and Superior Temporal Gyrus (LMTG), known as Wernicke's area. These will henceforth be referred to as Regions of Interest (ROIs). During sentence processing, these two areas showed a significant increase in activation when compared to their behavior during a control condition. It was also observed that brain activation went to a higher level, and remained at such a level for a longer period of time, during the processing of ambiguous sentences than it did during the processing of unambiguous sentences. As the processing of ambiguities leads to an increase in the demand for cognitive resources [3], such an increase results in additional cortical activity, which we are interested in characterizing.

1.3 Data Processing and Analysis

Each subject was presented a sequence of 20 trials (sentences), 10 ambiguous and 10 unambiguous, presented in a random order. Each trial consisted of the presentation of a sentence for 10 seconds, followed by a yes/no comprehension

question. Cortical activity during the processing of each sentence was recorded every 1500 msec, providing a time series that constituted that basis of our data. The part of the brain under scrutiny is divided into a number of volume elements called *voxels*, measuring $3.125x3.125x5$ mm. During the experiment, the *BOLD* (Blood Oxygen Level Dependent) signal at each voxel was measured every 1500 msec. This response is an indirect indicator of neural activity [2], and thus we will use the terms activity and amount of activation to refer to the level of the said response. An *image* is a set of such activation values, with one value per voxel. As images are acquired, the result is a succession of values for each voxel, containing its activation values at each instant of the experiment. The image are acquired not volumetrically but sequentially in slice planes 5 mm thick, with the acquisition of all 7 slices distributed over 1500 msec. The slight differences in acquisition time for the different slices is later corrected for by an interpolation technique.

The recordings available at each voxel will then consist of one time series of the activation for each sentence, as well some extra series corresponding to "baseline" activation during a control condition, during which the subject just fixates an asterisk instead of going through sentence-reading. Obviously, during the latter there should be no sentence-processing related activity.

The next step in the analysis was the identification of *active voxels*. These are voxels that display a significant activity in any of the experimental conditions. The activity during the experimental condition is gauged by comparing it with the one taking place during the control condition. This was done using a voxel-wise t-test comparing the activation level in the baseline condition and during all other conditions. A very high t threshold is used (equivalent to a Bonferroni correction for multiple comparisons) to identify voxels whose activation level during sentence processing differs significantly from their level during the control condition.

The time courses of the active voxels in a subject were then averaged across the 10 sentences of the same type and normalized as "percentage of activation above baseline". Afterwards, these were averaged across subjects. The end result was an average timecourse for every sentence trial, from which was obtained the average timecourse for each type of sentence. From analyses of variance of the data it was possible to conclude that brain activation went to a higher level, and remained at such a level for a longer period of time, during the processing of ambiguous sentences than it did during the processing of unambiguous sentences.

The expected model for this sentence processing task features each ROI recruiting voxels from a certain pool as resources for sentence processing in general. If a particular sentence demands more resources than are available in the pool (through its being ambiguous, for instance), more activity will be demanded from pool voxels and, eventually, other voxels might be recruited. We would like to extend this study by analyzing the degree and manner of involvement of these specially recruited voxels in the task being performed, and this effort will be described in Sect. 2.

2 Identifying Voxels with Varying Behaviour across Conditions

2.1 Our Approach and Related Work

The problem as we see it consists in ascertaining whether the behaviour of the BOLD response at some voxels is distinguishable between the two experimental conditions. If so, we would have candidate locations that might be supporting the additional activation required for processing ambiguous sentences, which could then be examined.

Currently, this question is addressed by identifying the most active voxels in two different experimental conditions, averaging their time series under each condition and then comparing the two averages. Identification of active voxels is done through a voxelwise t-test on one of the following: the difference between the mean activity during experimental and control conditions or the correlation of the voxel time series with a paradigmatic time series which corresponds to the expected response for voxels involved in the task.

A more agnostic approach is the clustering of the time series of all the voxels, guided either by known constraints on present cognitive processes or just using hierarchical clustering and using one of several possible metrics (see [7]). The centroids of the clusters thus found are then examined in the same way as the average time courses found through t-tests.

Yet another approach is to use a bayesian model of the fMRI signal at each voxel (see [4]). This can be used for questions beyond that of whether a given voxel is active or not, such as the influence of experimental condition on the parameters of the model, while being subject to assumptions regarding plausible signal shapes, noise and other factors.

A closer look at the t-test and clustering approaches will reveal that they give no guarantees of identifying voxels where the time series are different across experimental conditions. To see why consider that a high t-value for the contrast between experimental conditions and control only pertains to the mean activity and says nothing about the shape of the time series, which may be different for voxels that behave differently across the two experimental conditions In addition, the mean activity may be lower for such voxels than for the majority of voxels that accounts for the bulk of the activation.

If we were to user the t-test approach to test the mean during one experimental condition against the mean during another we would probably find that the means were too similar for strong results in most voxels where there is activity in both conditions. A clustering approach applied separately to each condition would still identify groups corresponding to the bulk of activation, which coincides in the two conditions.

The bayesian modelling approach allows for greater flexibility in that questions besides that of whether a voxel is active or not can be posed. In our case, the question would be whether the shape of the time series differs on a point by point basis across the two conditions. The caveat in this case is that the model pressuposes a certain BOLD response shape. Model fitting is accomplished by

estimating the values of the model parameters from data. While this is fine for the bulk of the active voxels, which mainly share the same response shape, it might not work for voxels where the shapes of the response are unusual in one or both conditions. Moreover, prior information about response shape is, in many cases, derived from observations in areas such as motor cortex, which need not be exactly the same for areas performing cognitive functions such as language processing.

Therefore, we would like to find differences in time series for a given voxel in a way that is independent both of assumptions regarding response shape and of considerations about level of activity. We propose to use a classifier to learn the difference, through identification of the two types of sentences under consideration, ambiguous and unambiguous with two classes of examples to be learned, on a voxel by voxel basis. The features on which the learning will be based are the activation values recorded for each voxel at each time point during the processing of a sentence. The examples are the time series for each sentence for both conditions.

There are other possibilities for representation of the time courses. In the extreme, one might just consider the mean activity. Another possibility is to represent the time series as a number of adjacent temporal sections, in terms of which the activity is described, which is what is done for the bayesian modelling approach cited above. Yet another would be to obtain derived features such as spectra obtained via fourier or wavelet transforms and cast the learning problem in terms of them.

Our hope is that the degree of success of a classifier on a voxel can be taken as an estimate of how much the activity in the corresponding voxel differs between ambiguous and unambiguous sentences, the two experimental conditions. This, in turn, should be an indication of the degree of involvement of the voxel in specifically processing ambiguity.

2.2 Experimental Procedure

The data available for each voxel consists of 10 sentences per condition, where each sentence is a time series of 16 activation values (24 sec of data) captured during and immediately after the processing of a sentence. We used only values 4 to 13 in each token, eliminating the pre-rise and post-activity decay of the BOLD response. Each of the values in a token has been normalized as a percentage, referring to how much it was above the average base level of activation during the control condition of the experiment.

In classifier terms, this maps to 10 examples of each class, where each example is a series of 16 floating point values, available for every voxel in both ROIs for 6 subjects.

In addition, given the already small number of examples we would like to have as few features per example as possible, and thus 10 features were retained in the part of the series where differences are more likely, as mentioned above.

As what is needed is an estimate of how accurately the difference between conditions at a given voxel can be learned using a given classifier, we resort to

leave-1-out cross validation over the 20 examples available, while taking care to balance the number of training examples across both classes. Initially, we create 10 random pairs with one example of each class. For each pair, we train on the remainder 18 examples, 9 of each class, and then test each of the examples of the pair. The estimate for attainable classifier accuracy for this voxel will be $\frac{\#ofsuccesses}{20}$.

The classifiers used in each voxel were a neural network (NN) with one sigmoidal hidden unit (see [6]), effectively doing logistic regression, and a linear kernel support vector machine (SVM) (see [5]). The choice of classifier was limited by the small number of examples available. We also tried other alternatives, such as SVMs with more complex kernels, NN with more hidden units and a simple k-nearest neighbour classifier. These were discarded because the performance was worse.

There are thus twelve sets of classification problems, with the set per each subject/ROI combination containing hundreds of classification problems, one per voxel. The output of the process is, for each subject and ROI, a list of voxels in that ROI ranked by decreasing classifier accuracy. Within groups of voxels with the same accuracy level, an additional ranking is performed based on "quality" of the classification (lowest mean squared error for NN or highest absolute value of the decision function for SVM). For each classifier, the process is run with different seeds for a number of times, the resulting rankings are averaged, and it is on these averaged lists that the analysis detailed below is performed.

2.3 Experimental Results

We found that it is possible to discriminate between activity in each of two experimental conditions for a small subset of the examined voxels. After applying the procedure described, we obtained consistent results across subjects and ROIs, in that a subset of 1% of the classifiers almost always has mean accuracy of 80% or more, for the best performing method (NN). The value of 1% typically corresponds to 3 to 7 voxels per subject/ROI (i.e. about 150 to 350 cubic mm) out of an anatomically huge volume of cortex. Considerations of how plausible the voxels found are from the psychological point of view are dealt with in the next section.

In each subject/ROI pair the distribution of accuracies is a unimodal curve centered around 50%, with a heavier tail for the higher accuracies and a smaller one for low accuracy voxels. Details about the distribution of the accuracy scores are discussed in the section that compares the results to a null model.

Because the goal was to find the relatively small group of voxels which activated differently during the processing of the two types of sentences, we will proceed considering only the top 1% most accurate voxels in each ROI and subject combination. Note that this number of voxels (3-7) is on the order of magnitude of the number of voxels believed to show real activation, typically demonstrated by their time-locking to the stimulus events and their signal amplitude in the experimental conditions, as detected through a t-test.

Table 1 details the mean accuracy attained in the top 1% group of voxels classified through each of the two methods. In addition, we were interested in finding out whether the voxels for which accuracy was greater using each method were the same, and thus the table contains the percentage of overlap between the groups. The comparison is made for the 6 subjects and 2 ROIs.

Table 1. Mean accuracy of the top 1% group selected using each classification method, as well as percentage of those belonging to the two groups

Subject#/ROI	**Accuracy NN**	**Accuracy SVM**	**Overlap %**
1-LIFG	0.82	0.80	0.50
1-LSTG	0.82	0.79	0.29
2-LIFG	0.78	0.72	0.00
2-LSTG	0.82	0.70	0.17
3-LIFG	0.82	0.81	0.25
3-LSTG	0.77	0.75	0.33
4-LIFG	0.85	0.78	0.67
4-LSTG	0.81	0.79	0.00
5-LIFG	0.86	0.76	0.00
5-LSTG	0.81	0.76	0.00
6-LIFG	0.81	0.83	0.33
6-LSTG	0.87	0.75	0.40
mean	0.82	0.77	0.25

The NN group had a mean accuracy ranging from 77% to 87% over subjects and ROIs, with an overall average of 82%, as displayed in the second column of Table 1. This means that a NN was capable of correctly distinguishing whether the subject had been reading an ambiguous or an unambiguous sentence on the basis of the fMRI activation data 82% of the time, on sentences on which it had not been trained, for each of these voxels. The SVM group had a mean accuracy ranging from 70% to 83%, averaging 77%. As this indicated some systematic difference between the two groups, we ran an ANOVA test on the difference in mean accuracy on the top 1% voxels between the two classifiers for all subjects/ROIs. In this test NN performed reliably better than SVM, across subjects and ROIs, at the 5% significance level, and therefore we will focus the rest of the paper on NN selected voxels.

Interestingly, there was not a great deal of overlap between the groups of most predictive voxels selected using each of the two methods, as evident in the mean overlap of 25%. This was unexpected, given that both methods resort to linear decision boundaries.

2.4 Comparison to a Null Model

As the classifiers for a few of the voxels considered do attain high levels of accuracy, we would like to demonstrate that that does not happen by chance.

In addition, we'd like to know for how many voxels can the effect be considered reliable.

Given the number of voxels involved and the procedure used, a classifier that guessed at random could conceivably attain relatively high accuracies in some small number of voxels. On the other hand, if we observed this happening for a large enough number of voxels, it would be less and less likely.

One way of testing this is to postulate a null model in which every voxel has the same inherent classifiability, and then examine what the probability of obtaining our accuracy results under that model is. By its being small it will be shown that the underlying model is not correct and that the inherent predictability at different voxels varies and can be high for a small group of them.

For a given voxel the accuracy of a classifier is an estimate of an underlying "true" accuracy attainable with that classifier. Given that the classifier is tested over 20 examples, the outcome can be seen as a sample from a binomial variable with 20 trials and probability of success equal to the underlying accuracy. In practice, the 20 trials are not independent, as they are a sequence of leave-1-out trials in which every pair of trials shares all but one training example. As performing the analysis in the latter case is far more complicated and would introduce details specific to the classification method used, we will proceed assuming independence. An empirical argument as to why the results thus obtained can still be used is given later.

Let us assume as a null hypothesis that the accuracy in each voxel is the same, and is some value close to 50%. The latter can be the empirical average of accuracies attained in most ROIs, which is indeed around that value.

We will examine the probability of the higher scores in a ROI under this model, assuming the ROI has n voxels:

$Pr(1^{st}max \geq a) = 1 - Pr(max < a) = 1 - (cdf(a))^n$

and, in general,

$Pr(k^{th}max \geq a) = \Pi_{i=1}^{k}(1 - (cdf(a))^{n-i+1})$

where $cdf(a) = Pr(X < a)$ when X $Binomial(20, \hat{p})$, with $\hat{p}$ being the mean observed accuracy across the ROI.

Under this model, we can calculate the probability that the k^{th} high score would be observed, and thus declare the probability as significant (and unexpected under the model) if it falls belows a certain threshold. For our analysis we used this criterion, and considered an accuracy level significant if had a probability of 5% or less of occurring under the model.

Table 2 contains the results of this experiment discriminated by subject/ROI combination, with the number of voxels considered significant out of the total, the least accuracy attained on one of those voxels and the percentage of voxels out of the total that constitute the group.

As stated before, these results are related to a null model where we assume that the results of each of the 20 leave-1-out trials for a voxel are independent, for simplicity reasons. Treating the case where the trials are not independent is,

Table 2. Breakdown of the number of voxels with significant differences across conditions

Subject#ROI	#significant	out of	accuracy of lowest	top %
1-LIFG	2	395	90	0.5
1-LSTG	0	617	0	0.0
2-LIFG	0	257	0	0.0
2-LSTG	3	543	85	0.6
3-LIFG	3	329	85	0.9
3-LSTG	1	536	95	0.2
4-LIFG	3	257	85	1.2
4-LSTG	0	516	0	0.0
5-LIFG	3	244	85	1.2
5-LSTG	2	204	85	1.0
6-LIFG	1	269	90	0.4
6-LSTG	5	415	85	1.2

in our view, too complicated, so instead we decided to run an empirical test, as follows.

The same setup as for the NN experiments was used, but the data was randomized. Five of the ten examples in each class were selected at random and switched to the other class. In this fashion each classifier was guaranteed to have five correct examples in each class and five incorrect ones, and its expected accuracy should not be more than 50%. Note that this is a case where the results in each of the 20 leave-1-out trials are certainly not independent, as each pair of trials shares most of the training data. The same number of repetitions were made and the results were ranked in the same way as originally.

Looking at the number of voxels given by the null model procedure in the randomized results we notice that their accuracies are far below what could be attained by the null model outputting random classifications.

The point is that the few classification results on the table are deemed improbable under the null model with the independence assumption, and they are far more improbable under a true model where the expected accuracy is 50% and where maximum accuracy practically never rises above 70%. As a consequence, it is reasonable to think that the test based on the independence assumption is conservative.

Moreover, most of the subject/ROI combinations contain a few voxels that are significant by this criterion. Given that our wish is to narrow down the number of voxels to be manually examined, we feel that the possibility of allowing some false positives in the group is acceptable. Nevertheless, we will proceed by considering only the voxels deemed significant through the procedure above.

Our conclusion is that for most voxels there is no inherent discernibility, but that it very probably exists for at least a small subset of them, and that this warrants the exploration described in the previous sections.

2.5 Voxel Characteristics

The group of voxels in consideration in each subject/ROI was picked because the activity in each voxel was discernible across the two experimental conditions. This may have been so because of heightened activity during the processing of ambiguous sentences. However, it may also mean that what stands out is the level of activity during the processing of unambiguous sentences as being unusual.

There are also no guarantees that the identified voxels have a high degree of activation, at least on average. In fact, for each combination of subject and ROI, the group identified almost always has no voxels in common with the subset of the top 1% most active voxels for the same combination, as selected through a t-test. This t-test compares the mean activity during experimental conditions and during a control condition which acts as a baseline.

One possible application of being able to find this group of voxels is to identify common characteristics in their time courses, such as the onset and duration of higher activity in a given experimental condition.

There was no clear trend in the logistic regression coefficients found, which was our initial expectation and would allow us to target a specific temporal region as the source of the difference. In addition, some of the voxels found displayed higher activity for ambiguous sentences, others for unambiguous ones.

The voxels with higher activity in ambiguous sentences did correspond to our expectation of the type of voxel recruited when resource demands are extreme. This is reflected in their not showing a consistent higher activation throughout the task, but rather high activity intervention in short bursts, which presumably would be where resources are more demanded for the ambiguous sentences.

A tentative explanation for this would be that different sentences are not matched for length, and thus the ambiguity, and the corresponding demand for extra resources, can occur in different points in time. This contrasts with the voxels in the most active group, which display consistent high activity throughout most of the of the task, as they are involved in the main processing, and higher activity during processing of ambiguous sentences.

A serious objection may be that many of the voxels selected do not show a clear spatial distribution, contrary to what happens for the active voxels, which tend to cluster tightly and where activation spreads radially as demand increases. While a few are adjacent to such active voxel groups, most are set at locations further away (still inside the ROI). If we expect a model where activity percolates to neighbours from a centre as demand rises, then this is hard to explain.

There were not significant contrasts in accuracy across subjects, ROIs and experimental conditions (across sentences being tested).

2.6 Conclusion and Further Work

We have presented a novel method for identifying voxels with contrasting behaviors across experimental conditions in a fMRI study. Through its use it was possible to find a subset of such voxels in a real dataset.

Unfortunately, further analysis of this subset of voxels across subjects failed to reveal any striking temporal patterns of activity or contrasts in accuracy

related to varying experimental variables (condition, subject, ROI). Many of the selected predicting voxels seem to have been picked for reasons not readily observable with the naked eye or easily relatable to the actual processing, which was our original hope. We are uncertain about what attributes of the time series of the predicting voxels are used by the classifier.

We do think, however, that the use of this method may be more successful in somewhat different studies. This would require alterations in experimental design so as to have sentences with similar lengths and positioning of ambiguity. Other possibilities lie in the use of alternative representations of the time series, be it through the use of composite features built from the initial series or through representations incorporating some prior assumptions regarding BOLD response shape. While limiting what can be learned, the latter might not be so restrictive as a full model of the response shape with parameters fitted to the data.

Another possibility would be to use the method to compare not two experimental conditions but all the conditions against the control condition. This could be used in place of t-tests for identifying active voxels as those where there is a greater contrast between activity during the experiment and activity during control, where the activity should be minimal and, above all, unstructured. The t-tests often performed for this purpose take into account solely the mean activity during experiment, and therefore our method would incorporate more information and possibly provide a better result. A completely different avenue rooted in the same idea would be the use of statistical tests of the difference between the distributions of time courses from the two classes for a given voxel (see, for instance, [9]).

Acknowledgments. Francisco Pereira was funded through a PhD scholarship from Fundação para a Ciência e Tecnologia, Portugal, for whose support he is exceedingly grateful. He would also like to thank Tom Minka and the anonymous reviewers for their precious comments and suggestions.

References

1. "Images of Mind", Posner, M.I, Raichle, M.E., W H Freeman (1997)
2. "Functional MRI", Moonen, C.T.W., Bandettini, P.A. (Eds.), Springer Verlag (2000)
3. "Ambiguity in the brain: How syntactically ambiguous sentences are processed.", Mason, R. A., Just, M. A., Keller, T. A., Carpenter, P. A., Manuscript submitted for publication (2001)
4. "Statistical Inference in Functional Magnetic Resonance Imaging", Genovese, C.R., Technical Report 674 (1999), Statistics Department, Carnegie Mellon University
5. "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning", Joachims, T., B. Schölkopf and C. Burges and A. Smola (ed.) MIT Press (1999)
6. "Machine Learning", Mitchell, T., McGraw Hill (1997)
7. "On clustering fMRI time series", Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., and Hansen, L. K., NeuroImage, 9(3):298-310.

8. “A neural network classifier for cerebral perfusion imaging”, Chan KH, Johnson KA, Becker JA, Satlin A, Mendelson J, Garada B, Holman BL., Journal of Nuclear Medicine 35(5):771-4. (1994)
9. “Discrete Multivariate Analysis: Theory and Practice”, Bishop, Y.M., Fienberg, S.E., Holland, P.W., MIT Press (1975)