Computational Cognitive Science (2017–2018)

School of Informatics, University of Edinburgh
Original exercises by Frank Keller, with modifications by Chris Lucas

Solutions for Tutorial 5: Word Recognition

Please work through this tutorial sheet on your own time as much as possible before arriving in tutorial. We encourage you to work together and discuss your methods and solutions.

Word Recognition and Neighborhoods

Question 1: In the lectures, we discussed a model of visual word recognition called Bayesian Reader proposed by Norris (2006). In this model, a visual input I is recognized as word $\hat{W}$ by computing:

$$\hat{W} = \arg\max_{W_i} P(W_i|I) = \frac{P(I|W_i)P(W_i)}{P(I)} \tag{1}$$

where W_i ranges over all words in the lexicon.

1. Which factors influence word recognition according to equation (1)? Explain the role of each of the terms on the right-hand side of the equation.

2. Lexical decision is the task of deciding for a given input (typically a string of letters) whether it is a word of the language or not. Show how Bayesian Reader can model lexical decision by replacing the MAP computation in equation (1) with a Bayesian summation.

3. Comment on the cognitive plausibility of the Bayesian computation in the previous question. Which additional assumptions may be necessary to increase the cognitive plausibility of this aspect of the model?

4. Is the issue discussed in the previous question a general feature of Bayesian models? Consider the other Bayesian models discussed in the course so far: The Tenenbaum model of concept learning, and the Frank et al. model of word learning.

Solution 1:

1. The factors that influence word recognition time and accuracy are:
   (a) the probability of the input: $P(I)$ represents how probable a given input string is; it can be reduced if the input is blurred, distorted (handwriting etc.), or lighting is poor;
   (b) the prior probability of the word: $P(W_i)$ depends on the frequency of W_i and its plausibility in a given context;
   (c) how likely a word is to generate the input: $P(I|W_i)$ depends on how much time has passed since the word was presented, i.e., how many samples the visual system was able to take; the more samples, the more certain we are that it is W_i that gives rise to I.

2. We are now interested in the probability that the input is a word, which we encode using the binary random variable *word*. Instead of computing $\hat{W}$, we now compute $P(word|I)$:

$$P(word|I) = \frac{P(I|word)P(word)}{P(I)} \propto P(I|word)P(word)$$

we can assume that $P(word)$ is constant, as a given experiment will contain a constant proportion of words and non-words (e.g., 50% each). And we can work out the likelihood $P(I|word)$ by summing over all possible words W_i:

$$P(I|word) = \sum_i P(I|word, W_i)P(W_i|word) = \sum_i P(I|W_i)P(W_i) \qquad (2)$$

This computation is Bayesian, because it involves summing over hypotheses (words in this case).

3. The Bayesian computation involves summing over all words in the lexicon. This is a large sum (tens, maybe hundreds of thousands of words), which has to happen in the real time, in the space of a few hundred milliseconds. It's debatable whether such a computation is cognitively plausible. Additional assumptions are necessary, for example that the summation only considers words of the same length (if we assume that word length can be recognized without performing word recognition proper).
4. Bayesian summation is a feature of all Bayesian models. For example, the Tenenbaum model involves summing over all possible concepts (hypotheses), and the Frank et al. model involves summing over all referents and all objects. However, these models do not model a real-time process. Concept learning presumably involves lengthy deliberation, and lexicon acquisition happens over a long period of time (perhaps with consolidation during sleep). So in this sense, the problem is specific to Bayesian Reader.

Question 2: Experiments show that the time it takes to recognize a written word depends on the number of *orthographic neighbors* of the word to be recognized. An orthographic neighbor is defined as a word of the same length that differs from the original string by only one letter.

1. Draw a neighborhood graph for the following words. In such a graph, the words are nodes, and two nodes are connected by an edge if the words represented by the edges are orthographic neighbors.

   cat, bat, fat, cab, bar, bet, but, bus

   What is the maximum number of neighbors a word of three letters can have in principle?

2. Figure 1 plots the number of orthographical neighbors against the reaction time in a lexical decision experiment. Explain the effects that the graph shows. How does Bayesian Reader model them?

3. How would you expect the neighborhood effects shown in Figure 1 to change in non-native speakers? How could you model this with Bayesian Reader?

Solution 2:

1. This should be easy.
2. The figure graphs the neighborhood effect for words and non-words: words show decreased reaction time with increasing number of neighbors; non-words show increased reaction time. Equation (2) explains this as follows: if a word has many neighbors, then many words will have a high $P(I|W_i)$. This results in a higher likelihood $P(I|word)$, so word decisions speed up with increasing neighborhood size. This applies equally to non-words (note that the neighbors of a non-word are words, not other non-words): if a non-word has many neighbors, then $P(I|word)$ will be higher, hence $P(I|non\text{-}word)$ will be smaller relative to it, so non-word decisions are slower.

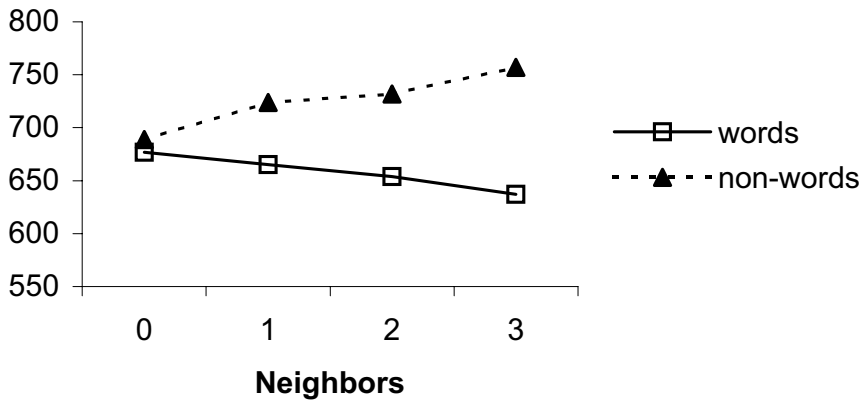

Figure 1: Reaction time vs. number of neighbors in a lexical decision experiment

3. This is pretty open-ended question. The answer depends on what you what to assume about non-native speakers and their lexicon. For example, presumably they have been exposed to less text and speech in their non-native language, so the prior based on word frequencies would be smaller.

   It's also safe to assume that the lexicon of non-native speakers is smaller, resulting in less neighbors overall, so lexical decision times for words will be slower. This would affect the likelihood term of Bayesian Reader.

   Things get more interesting if we consider a bilingual lexicon – non-native speakers of course need to represent the words of their native language too. The two lexicons are not completely separate (we know this from crosslinguistic priming experiments). So for example, the English word *cat* could then have French neighbors as well as English neighbors. Probably these two types of neighbors should have affect the likelihood in different ways.

References

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.