Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 and MoSIG M1 Winter Semester 2017

Exercise 8 31 March 2017

The goal of this exercise is to write CLIPS rules that will make it possible to recognize categories of text from word N-Grams. An N-Gram is a sequence of N symbols. N-grams of words are a common feature used to classify text. In this exercise we will use clips rules to count the frequency of occurrence of word 2-Grams (word-pairs) in different categories of text. Example categories can include email, SMS, technical writing, publicity, spam, etc.

You can assume that text is provided in the form of paragraphs. A function named "read-paragraph" is provided to creates a fact in working memory of the form:

```
(Paragraph class  w1 w2 … wN)
```

where the <wn> are the individual words of the paragraph.

Your system should use the following templates for Word-Pair and Category.

```
(deftemplate WordPair   ; structure for ccounting Word Pairs (2-Grams of words)
      (slot CATEGORY (type SYMBOL))
      (slot WORD1 (type SYMBOL))
      (slot WORD2 (type SYMBOL))
      (slot COUNT (type INTEGER))  ; Number of instances of word pair
)

(deftemplate Category   ; A category of text (e.g. scientific, legal, spam, etc)
      (slot NAME) (type SYMBOL)) ; Name for Category of Text
      (slot M (type INTEGER)) ; Size of Training Set for Category
```

a) Write a rule named MakeWordPair to generate a fact of type WordPair for each new word pair in a paragraph of a sample of a category. Be sure to include the category when you create each Word-Pair. Be sure not to create more than one fact for each word pair.

b) Write a rule named CountWordPairs that updates the count for each word-pair in the paragraph. Make sure that this rule also updates the size of the training set, M, for the category.

c) Assume that you have a very large corpus (training set) of text that includes several paragraphs of several categories. Write a rule or set of rules to determine the sum of the counts for each word-pair for all categories. Store the result as facts of Word-Pair with a category name of "All". Name your rule(s) CountPairsForAllCategories

d) Write a rule to print the most frequent word-pair in each category, along with its count. If several word-pairs have the same most frequent count, then print them all. The printed message should say:

"The most frequent word pair in category <C> is <W1> <W2> with count <N>"

where <C>, <W1>, <W2>, <N> represent the category, words and count.