Problem 1: BIO Annotation Scores

$$\boxed{Kappa\ Statistic:\ \kappa = \frac{P(agree) - P(expected)}{1 - P(expected)}} \tag{1}$$

$$\boxed{P(expected) = \sum_{c \in C} P(c|A_1) * P(c|A_2)} \tag{2}$$

(a) Using the two sets of annotations we can calculate the conditional probabilities needed to calculate κ.

Probability	Value
$P(B\|A_1)$	6/24
$P(I\|A_1)$	8/24
$P(O\|A_1)$	10/24
$P(B\|A_2)$	8/24
$P(I\|A_2)$	4/24
$P(O\|A_2)$	12/24

Then we can calculate $P(agree)$ and $P(expected)$.

$$P(agree) = 15/24$$

$$P(expected) = (6/24 * 8/24) + (8/24 * 4/24) + (10/24 * 12/24) = 25/72$$

Finally we compute the Kappa Statistic.

$$\kappa = \frac{(15/24) - (25/72)}{1 - (25/72)} = 20/47 = 0.4255$$

$$\boxed{\kappa = 0.4255}$$

$$\boxed{Recall:\ \frac{\#\ correctly\ labeled\ as\ C}{\#\ true\ instances\ of\ C}} \tag{3}$$

$$\boxed{Precision:\ \frac{\#\ correctly\ labeled\ as\ C}{\#\ labeled\ as\ C}} \tag{4}$$

(b) Assuming that A_1 annotations were produced by a human and A_2 by an IE system, we can calculate the recall and precision of each of the 3 BIO labels.

$$Recall_B = \frac{\#\ correctly\ labeled\ as\ B}{\#\ true\ instances\ of\ B} = \boxed{3/6}$$

$$Precision_B = \frac{\#\ correctly\ labeled\ as\ B}{\#\ labeled\ as\ B} = \boxed{3/8}$$

$$Recall_I = \frac{\#\ correctly\ labeled\ as\ I}{\#\ true\ instances\ of\ I} = \boxed{4/8}$$

$$Precision_I = \frac{\#\ correctly\ labeled\ as\ I}{\#\ labeled\ as\ I} = \boxed{4/4}$$

$$Recall_O = \frac{\#\ correctly\ labeled\ as\ O}{\#\ true\ instances\ of\ O} = \boxed{8/10}$$

$$Precision_O = \frac{\#\ correctly\ labeled\ as\ O}{\#\ labeled\ as\ O} = \boxed{8/12}$$

(c) Assuming that A_2 annotations were produced by a human and A_1 by an IE system, we can calculate the recall and precision of each of the 3 BIO labels.

$$Recall_B = \frac{\#\ correctly\ labeled\ as\ B}{\#\ true\ instances\ of\ B} = \boxed{3/8}$$

$$Precision_B = \frac{\#\ correctly\ labeled\ as\ B}{\#\ labeled\ as\ B} = \boxed{3/6}$$

$$Recall_I = \frac{\#\ correctly\ labeled\ as\ I}{\#\ true\ instances\ of\ I} = \boxed{4/4}$$

$$Precision_I = \frac{\#\ correctly\ labeled\ as\ I}{\#\ labeled\ as\ I} = \boxed{4/8}$$

$$Recall_O = \frac{\#\ correctly\ labeled\ as\ O}{\#\ true\ instances\ of\ O} = \boxed{8/12}$$

$$Precision_O = \frac{\#\ correctly\ labeled\ as\ O}{\#\ labeled\ as\ O} = \boxed{8/10}$$

Problem 2: k-fold Cross-Validation

(a) 4,500 documents

(b) 27 documents

(c) 4,000 documents

(d) 24 documents

(e) If I am only allowed to perform one cross-validation experiment, I would choose the **astronomy corpus**. When performing cross-validation, all of the available data is used both for training and testing. This is accomplished by performing N experiments (where N is your fold size) allowing all of your data to be used for training or testing at some point during the N experiments. Since the size of the astronomy corpus is much smaller than the medical corpus, I feel like this is the right choice.

Problem 3: Hidden Markov Model

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) \tag{5}$$

Where $\alpha_{t-1}(i)$ is the **previous forward path probability** from the previous time step, a_{ij} is the **transition probability** from previous state q_i to the current state q_j, and $b_j(o_t)$ is the **state observation likelihood** of the observation symbol o_t given the current state j. [1]

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{6}$$

Where a_{ij} is the **transition probability** from from previous state q_i to the current state q_j, $b_j(o_{t+1})$ is the **state observation likelihood** of the observation symbol o_{t+1} given the current state j, and $\beta_{t+1}(j)$ is the **backward path probability** from the next time step. [1]

(a)

$$\alpha_1(LOC) = P(LOC|\phi) * P(Utah|LOC) = 0.40 * 0.08 = \boxed{0.032}$$

$$\alpha_1(ORG) = P(ORG|\phi) * P(Utah|ORG) = 0.25 * 0.05 = \boxed{0.0125}$$

$$\alpha_1(NONE) = P(NONE|\phi) * P(Utah|NONE) = 0.35 * 0.02 = \boxed{0.007}$$

$$\begin{aligned}\alpha_2(LOC) = &\alpha_1(LOC) * P(LOC|LOC) * P(Grizzlies|LOC) + \\ &\alpha_1(ORG) * P(LOC|ORG) * P(Grizzlies|LOC) + \\ &\alpha_1(NONE) * P(LOC|NONE) * P(Grizzlies|LOC) \ = \\ &0.032 * 0.6 * 0.03 + 0.0125 * 0.15 * 0.03 + 0.007 * 0.08 * 0.03 = \boxed{0.000649}\end{aligned}$$

$$\begin{aligned}\alpha_2(ORG) = &\alpha_1(LOC) * P(ORG|LOC) * P(Grizzlies|ORG) + \\ &\alpha_1(ORG) * P(ORG|ORG) * P(Grizzlies|ORG) + \\ &\alpha_1(NONE) * P(ORG|NONE) * P(Grizzlies|ORG) \ = \\ &0.032 * 0.1 * 0.06 + 0.0125 * 0.7 * 0.06 + 0.007 * 0.02 * 0.06 = \boxed{0.000725}\end{aligned}$$

$$\begin{aligned}\alpha_2(NONE) = &\alpha_1(LOC) * P(NONE|LOC) * P(Grizzlies|NONE) + \\ &\alpha_1(ORG) * P(NONE|ORG) * P(Grizzlies|NONE) + \\ &\alpha_1(NONE) * P(NONE|NONE) * P(Grizzlies|NONE) \ = \\ &0.032 * 0.3 * 0.01 + 0.0125 * 0.25 * 0.01 + 0.007 * 0.9 * 0.01 = \boxed{0.00019}\end{aligned}$$

(b)

$$\beta_3(LOC) = P(\Omega|LOC) * P(Win|LOC) = 0.5 * 0.04 = \boxed{0.02}$$

$$\beta_3(ORG) = P(\Omega|ORG) * P(Win|ORG) = 0.2 * 0.02 = \boxed{0.004}$$

$$\beta_3(NONE) = P(\Omega|NONE) * P(Win|NONE) = 0.3 * 0.09 = \boxed{0.027}$$

$$\begin{aligned}\beta_2(LOC) = &P(LOC|LOC) * P(Win|LOC) * \beta_3(LOC) + \\ &P(ORG|LOC) * P(Win|ORG) * \beta_3(ORG) + \\ &P(NONE|LOC) * P(Win|NONE) * \beta_3(NONE) \ = \\ &0.6 * 0.04 * 0.02 + 0.1 * 0.02 * 0.004 + 0.3 * 0.09 * 0.027 = \boxed{0.001217}\end{aligned}$$

$$\begin{aligned}\beta_2(ORG) = &P(ORG|LOC) * P(Win|LOC) * \beta_3(LOC) + \\ &P(ORG|ORG) * P(Win|ORG) * \beta_3(ORG) + \\ &P(NONE|ORG) * P(Win|NONE) * \beta_3(NONE) \ = \\ &0.1 * 0.04 * 0.02 + 0.7 * 0.02 * 0.004 + 0.25 * 0.09 * 0.027 = \boxed{0.000744}\end{aligned}$$

$$\begin{aligned}\beta_2(NONE) = & P(LOC|NONE) * P(Win|LOC) * \beta_3(LOC) + \\ & P(ORG|NONE) * P(Win|ORG) * \beta_3(ORG) + \\ & P(NONE|NONE) * P(Win|NONE) * \beta_3(NONE) \; = \\ & 0.08 * 0.04 * 0.02 + 0.02 * 0.02 * 0.004 + 0.9 * 0.09 * 0.027 = \boxed{0.00225}\end{aligned}$$

Problem 4: Model Choices

(a) ORDINARY CLASSIFIER - Spam detection is a binary classification task. We care about if the sentence as a whole is spam and not necessarily interested in classifying each word in the sentence. That said, we could still build features that detect keywords commonly found in spam and look use features such as the number of exclamation points in the sentence.

(b) SEQUENCE TAGGING - Identifying the names of colleges and universities would rely on the sequence of words. For example, the word "university" shortly followed by a capital word such as "Utah" would strongly indicate a university name.

(c) SEQUENCE TAGGING - Sequential information would be very useful in this task to determine if we are observing a list of cars. For example, looking at the previous and following words would be strong features for classification. Along with features about the words themselves.

(d) SEQUENCE CLASSIFIER - While this is a binary classification task, I think sentence structure will be a strong indicator of labeling. For example, a child would be more inclined to grammatical errors (in theory) and sentence structure would reveal this. Additionally features dealing with spelling errors could be worked into this model.

(e) ORDINARY CLASSIFIER - This is a binary classification task. We aren't trying to identify information about each word in the sentence, but rather classify the sentence as a whole. Features such as keywords could be engineered to classify if the topic is computer science or chemistry.

References

[1] Jurafsky, D. & Martin, J. H. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. (Pearson Prentice Hall, 2009).