Inferring a Noun Phrase Clustering for an Antecedent-based Co-reference Resolution Model

Jason D. M. Rennie
jrennie@csail.mit.edu

November 12, 2004

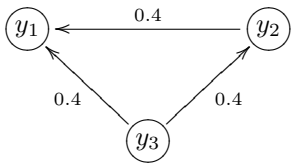

Figure 1: An example where a greedy top-down inference method would find a locally optimal labeling. Link annotations are antecedent probabilities. Not shown are probabilities of no antecedent. The greedy algorithm would assign different labels for the first two noun phrases, $y_1 = 1$ and $y_2 = 2$, $P_l(\vec{y}) = 0.24$; the optimal assignment, $P_l(\vec{y}) = 0.32$, is $y_1 = y_2 = y_3 = 1$.

We assume that a set of noun phrases $\{x_1, \ldots, x_n\}$ have been extracted from a text. We also assume the parameters for our model, $\vec{w}$, have been learned (Rennie, 2004). We would like to determine a configuration (set of labels), $\vec{y}$, that maximizes the joint likelihood of the model. Since our model is probabilistic, approximate inference[1] can be achieved via belief propagation (BP). However, since the underlying graph is fully connected, BP is not very useful. We instead consider a set of simpler, greedy algorithms.

The first algorithm we consider is also the simplest. We call it MaxAntecedent. Noun phrases are ordered according to their appearance in text. In order, a label is chosen for each noun phrase according to the maximum likelihood antecedent. That is, each noun phrase takes on the label of the noun phrase that has the highest antecedent probability.

The next algorithm is a variation on MaxAntecedent that is more in-line with our joint likelihood objective. We call it GreedyTopDown. Again, noun phrases are ordered according to their appearance in text. In order, labels are chosen to maximize conditional likelihood of the noun phrase label given the labels of all preceeding noun phrases. Note that if a noun phrase has many low-probability

[1]The underlying graph is fully connected, so we are not guaranteed to find the globally optimal configuration.

antecedents with the same label, it may choose that label over the label of a single noun phrase with high antecedent likelihood. Figure 1 shows a scenario for which this algorithm does not find the maximum likelihood configuration.

The next algorithm utilizes marginal label distributions. We call it Greedy-Marginal. The marginal label distribution for any noun phrase is simply a mixture of the marginal label distributions of the preceding noun phrases:

$$P_l(Y_i = y) = \sum_{j=1}^{i} e^{s(x_i, x_j)} P_l^j(y), \tag{1}$$

where $P_l^j(Y_i = y)$ a special distribution; if $j \neq i$ it is the usual marginal distribution, $P_l^j(Y_i = y) = P_l(Y_i = y)$; if $j = i$ then it is unity, $P_l^i(Y_i = y) = 1$. The inference algorithm procedes as follows. Until all noun phrases have been assigned a label, determine the noun phrase with the lowest entropy marginal label distribution. Assign that noun phrase the label with the maximum marginal probability. Update all other marginal distributions given the new evidence. Repeat.

To determine a labeling for the data, we run each of these algorithms; each yields a labeling of the data. We evaluate each labeling using the joint objective and choose the one with the greatest joint likelihood. We note that an issue with joint likelihood maximization for inference is that it has a tendency to assign the majority label.

References

Rennie, J. D. M. (2004). Learning parameters for an antecedent-based coreference resolution model. http://people.csail.mit.edu/∼jrennie/writing.