Learning Parameters for an Antecedent-based Co-reference Resolution Model

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

November 11, 2004

We assume that a set of noun phrases $\{x_1, \ldots, x_n\}$ have been extracted from a text. We assume that we are told the entity $\{E_1, \ldots, E_n\}$ to which each refers. We translate these entity names to integer labels by associating each entity with its first mention in the text. Define

$$y_i = \min\{j | E_j = E_i\}. \tag{1}$$

In other words, the label y_i for noun phrase x_i is the index of the first mention of the entity to which x_i refers.

We define a similarity function between two noun phrases[1]

$$s(x_i, x_j) = \vec{w} \cdot \vec{f}(x_i, x_j), \tag{2}$$

where $\vec{w}$ is a parameter vector and $\vec{f}$ is a vector of feature functions, each of which returns a real number depending on the pair of noun phrases. We assume the feature functions are fixed beforehand. We use training data to learn the parameter vector. Using the similarity function, we define an antecedent distribution on each noun phrase. The probability that x_j is the antecedent for x_i is proportional to the exponentiated similarity between the two noun phrases. We use the "self-antecedent" probability to represent the chance that the noun phrase has no antecedent.

$$P_a(x_i \to x_j) \propto \left\{ \begin{array}{cl} e^{s(x_i,x_j)} & \text{if } j \le i \\ 0 & \text{if } j > i \end{array} \right. . \tag{3}$$

A noun phrase refers to the same entity as its antecedent, so it inherits its antecedent's distribution over entity labels. Since there is a distribution over possible antecedents, the label distribution for a noun phrase becomes a mixture of experts with the mixture weights corresponding to antecedent probabilities. Given the labels of all preceeding noun phrases, the probability that a noun

[1] The similarity is defined for all possible noun phrase pairs, including $i = j$. When $i = j$, it takes on special meaning; it is not the similarity between the noun phrases, but rather a measure related to the chance that the noun phrase has no antecedent.

phrase takes on a certain label is the sum of antecedent probabilities for noun phrases with that label,

$$P_l(Y_i = y|y^{i-1}) = \frac{Q_i}{Z_i} = \frac{1}{Z_i}\sum_{j=1}^{i} e^{s(x_i,x_j)}\delta(y = y_j). \tag{4}$$

We use y^{i-1} to denote the labels of preceeding noun phrases, $\{y_1, \ldots, y_{i-1}\}$; $\delta(\cdot)$ is 1 if the condition is true, 0 otherwise. The product of these conditional distributions yields a joint distribution on labelings which correspond to partitionings of the data,

$$P_l(y^n) = \prod_{i=1}^{n} P_l(y_i|y^{i-1}). \tag{5}$$

We learn a weight vector by maximizing the difference between the joint log-likelihood and the squared L2 norm of the weight vector,

$$J = \log P_l(y^n) - \frac{C}{2}\|\vec{w}\|^2. \tag{6}$$

$$J = \sum_i \log Q_i - \sum_i \log Z_i - \frac{C}{2}\|\vec{w}\|^2. \tag{7}$$

We use a gradient descent-type algorithm for minimizing the objective. We find that the gradient fits the usual form of an exponential; the primary difference is that the empirical mean is an average of feature values of preceeding noun phrases in the same cluster,

$$\frac{\partial J}{\partial \vec{w}} = \sum_{i=1}^{n} \sum_{j \le i | y_j = y_i} \frac{\vec{f}(x_i,x_j)e^{s(x_i,x_j)}}{\sum_{j \le i | y_j = y_i} e^{s(x_i,x_j)}} - \sum_{i=1}^{n} \sum_{j \le i} \frac{\vec{f}(x_i,x_j)e^{s(x_i,x_j)}}{\sum_{j \le i} e^{s(x_i,x_j)}} - C\vec{w}. \tag{8}$$

Any number of gradient descent algorithms can be used to find a local maximum of the objective.

The objective is not convex. So, in order to improve our chances of finding a solution close to the global optimum, we use multiple starting points. We use the obvious starting point of $\vec{w} = \vec{0}$. We also use as a starting point the parameter vector that optimizes the model that assumes that the antecedent graph for each cluster forms a single chain. We use as an additional starting point the parameter vector that maximizes the McCallum and Wellner (2003) model[2].

References

McCallum, A., & Wellner, B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. *Proceedings of the IJCAI Workshop on Information Integration on the Web.*

[2] Note that our model may include weights for “self-antecedent” comparison that would make no sense for the McCallum and Wellner model. Those weights which would not be used in their model are set to zero.