A Max-Antecedent Model for Co-reference Resolution

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

November 14, 2004

We consider an alternative to our earlier model (Rennie, 2004). There, the entity distribution for a noun phrase was a mixture of experts where mixture weights were based on similarity and the set of experts was all preceding noun phrases. This creates an unfortunate majority bias. Even when there is no single high similarity preceding noun phrase for an entity, that entity may still be chosen; the entity probability is a sum over preceding noun phrases; sufficiently many low-similarity noun phrases can overcome a single high-similarity one. We avoid this bias by using the max instead of sum operator—the entity probability for a noun phrase is proportional to the maximum exponentiated similarity of any preceding noun phrase that refers to that entity. There is also a chance that the noun phrase does not have an antecedent—it refers to an entity that has yet to be mentioned. The new conditional probability is

$$P_l(Y_i = y|y^{i-1}) = \frac{1}{Z_i}\left(e^{(x_i,x_i)}\delta(y=i) + \max_{j<i|y_j=y}\left\{e^{s(x_i,x_j)}\right\}\right), \qquad (1)$$

where $Z_i = e^{(x_i,x_i)} + \sum_y \max_{j<i|y_j=y}\left\{e^{s(x_i,x_j)}\right\}$. The joint probability is a product of the conditionals: $P_l(\vec{y}) = \prod_i P_l(y_i|y^{i-1})$.

Note that the joint model is the max of a set of convex functions. We can achieve a local optimum via a basic gradient descent/line search algorithm. Some care must be taken with the line search since the function may not be convex over the domain of the search. A simple and effective technique is to maximize the the local, convex function at each iteration. Since the global objective is a maximum of these functions, an increase in the local objective is sure to increase the global objective.

Inference is hard. Note that determining a partitioning of the noun phrases is equivalent to determining the antecedent for each. The parameter vector tells us which preceding noun phrase maximizes the conditional probability. But, the antecedent choice has an impact on later conditional distributions. Later normalization constants are highly dependent on the configuration of earlier noun phrases. In particular, each new cluster dilutes later conditional distributions. However, the obvious top-down, greedy algorithm serves as a good approximate

inference method, especially if there are a very limited number of high-similarity possible antecedents per noun phrase. It also eliminates the bias against new clusters which is inherent in the joint objective.

References

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