A Hybrid Model for Co-reference Resolution

Jason Rennie

jrennie@csail.mit.edu

Joint work with Tommi Jaakkola

Claims & Contribution

- Proper noun resolution best handled via clustering
- Pronoun resolution best handled via classification
- Contribution:
  - A new classification model for non-proper nouns that integrates with McCallum and Wellner (2003) clustering model
- Sorry, no results yet :-(

Look here for notation definitions

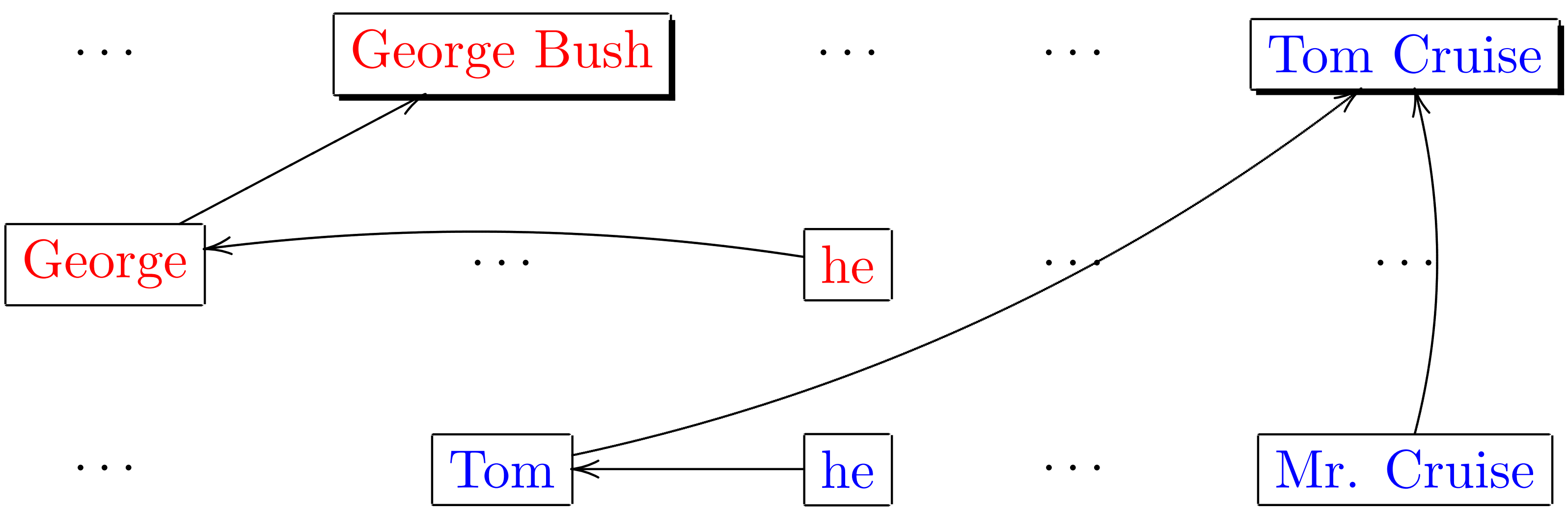
Co-reference Resolution

Goal is to group noun phrases according to entity reference:

. . .	George Bush	. . .	. . .	Tom Cruise
George	. . .	he	. . .	. . .
. . .	Tom	he	. . .	Mr. Cruise

Antecedent Structure

- Antecedent: a phrase or clause that is referred to by an anaphor
- Two NPs in same group if they are connected by antecedent links



Hidden Information

- Antecedent Structure is not given.
- Earlier work: assume a single chain of references, or assume that all within-cluster pairs have antecedent relation.
- Our approach: Learn the antecedent structure.

Multi-class Classification

- Order noun phrases: $\{x_1, \ldots, x_n\}$
- Assume: antecedent of each noun phrase must come before
- Probability that antecedent for x_i is x_j is log-linear:

$$P_a(x_i \rightarrow x_j) = \begin{cases} \frac{1}{Z_i} \exp(s(x_i, x_j)) & j < i \\ 0 & \text{othw.} \end{cases} \tag{1}$$

Similarity

- Similarity between two noun phrases as:

$$s(x_i, x_j) = \vec{\theta} \cdot \vec{f}(x_i, x_j) \qquad (2)$$

- Each *pair* of noun phrases define a feature vector
- Parameter vector determines how to combine features to create a similarity score
- Important: parameter vector is independent of number of clusters

A Forest

- What is the label of a noun phrase?
  - *Same label as it's antecedent!*
- Assume that proper noun phrases have been clustered
- A pronoun is grouped with a cluster if it has an antecedent chain to that cluster

Mixture of Experts

- Relaxed version: mixture of experts
- Antecedent probability for x_j is expert weight

$$P(Y_i = y|y^{i-1}) = \sum_{i<j} P_a(x_i \to x_j) P(Y_j = y|y^{j-1}) \qquad (3)$$

$y^k \equiv \{y_1, y_2, \ldots, y^k\}$

Model of McCallum and Wellner (2003)

- Use pairwise potentials to define a joint distribution (on proper noun phrases)

$$\psi(x_i, x_j, y_{ij}) = \exp(y_{ij} s(x_i, x_j)) \tag{4}$$

$$P(y^n) = \frac{1}{Z_{\vec{x}}} \prod_{i,j} \psi(x_i, x_j, y_{ij}) \tag{5}$$

- We use the same similarity function for both models (same parameters)

$$y_{ij} = \begin{cases} +1 & \text{if } y_i = y_j \\ -1 & \text{if } y_i \neq y_j \end{cases}$$

Hybrid Model

- McCallum and Wellner give joint model on proper noun phrases, $P(y_A)$
- We give conditional distribution on labels for each non-proper noun, $P(Y_i = y|y^{i-1})$
- Use Bayes' Law twice:
  - Product of conditionals yields non-proper noun model: $P(y_B|y_A) = \prod_{i \in B} P(y_i|y^{i-1})$
  - Full joint is product of two models: $P(y_A, y_B) = P(y_A)P(y_B|y_A)$

A: set of proper noun phrases

B: set of non-proper noun phrases

Learning & Inference

- Simple approach: maximize joint likelihood: $\max P(y_A, y_B)$
  - Unlike earlier work, antecedent information is recovered, clear how training data should be used
- Better approach: maximize product of marginals: $\max P(y_1)P(y_2)\cdots P(y_n)$ (Kakade et al., 2002)
  - Marginal objective better approximates zero-one error
  - But, more computationally difficult: each marginal is sum over joint, $P(y_1) = \sum_{\vec{y}|y_1} P(y_1, \ldots, y_n)$

Learning: maximize over θ; Inference: maximize over $\vec{y}$

Learning & Inference

- Learning determines parameters of similarity function
  - Learning also recovers (non-proper) antecedent distribution
  - Use Gradient Descent or Expectation-Maximization
- Inference is computationally difficult
  - McCallum and Wellner model is non-convex; our model adds additional non-convexities (antecedent graph may depend on proper noun partitioning)

Summary & Questions

- Classification model on antecedent relations well-suited for non-proper noun resolution
- Model integrates nicely with McCallum and Wellner (2003) clustering model (proper nouns)
- What to do with non-proper non-pronouns? Which is better model?
- How to make inference efficient without severe approximations?

References

Kakade, S., Teh, Y. W., & Roweis, S. (2002). An alternate objective function for markovian fields. *Proceedings of the Nineteenth International Conference on Machine Learning.*

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.*