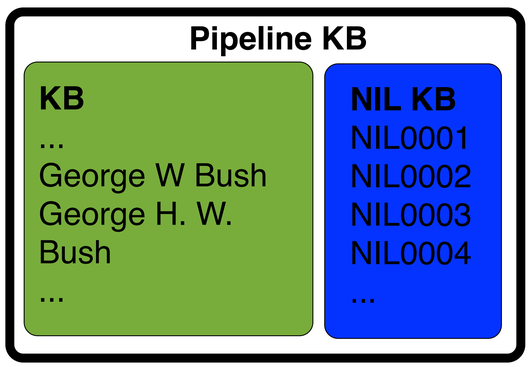
Cross-Document Coreference Resolution and Entity Linking using a Dirichlet Process

Travis Wolfe, Nicholas Andrews, Matt Gormley, Mark Dredze
Johns Hopkins University, Baltimore MD
{travis,noa,mrg,mdredze}@cs.jhu.edu

human language technology center of excellence

Problem

- Link mentions in text to entities in a knowledge base, or NIL if the mention does not refer to an entity in the KB
- Previous work handles NILs in an ad hoc fashion
- Goals
  1) run with little or no supervision
  2) do entity linking and NIL disambiguation jointly



Previous work classifies and clusters entities in a pipeline.

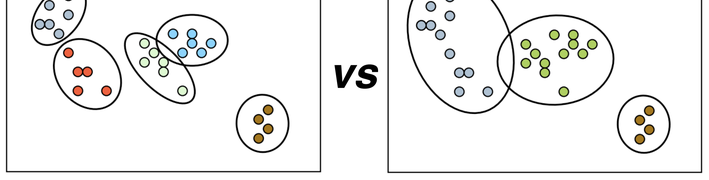
Our work jointly clusters known entities and NILs.

This Work

We propose a new generative model that jointly links mentions to a knowledge base and clusters NIL mentions with little dependency on supervision. Our model disambiguates mentions based on context words, name similarity, and popularity.

Dirichlet Process Prior

- We put a Dirichlet Process prior on clusters of mentions
- This is a non-parametric prior that allows the number of clusters to vary with the amount of data observed
- The correct DP prior allows the model to perform well without knowing how many entities exist.



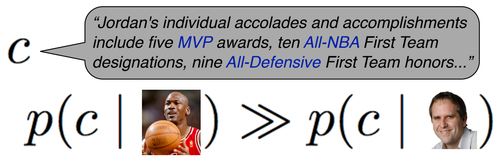
Name Model

- Our model assumes that there is a "canonical name" for every entity
- Mentions commonly drop middle names, abbreviate full names, use initials, add titles, etc.
- Learn a weighted FST to describe p(mention | name)

$$p(\text{“Thomas Jefferson”} \rightarrow \text{“Mr. Jefferson”}) = 0.18$$
$$p(\text{“Thomas Jefferson”} \rightarrow \text{“Mr. Thomas”}) = 0.034$$

Context Model

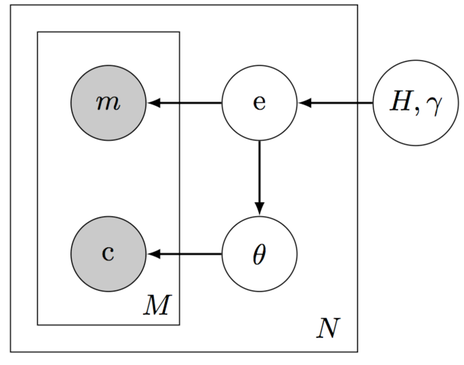
- Context of a mention can help disambiguate entities
- Our generative model includes a Dirichlet-Multinomial language model for context words.



Inference

- Gibbs sampling for inference
- Sample cluster membership for mentions according to the equation at the bottom
- Since this is a very costly distribution to normalize (there are as many clusters as there are entities, which can be millions), we approximate this distribution by assuming that only a few clusters have non-zero mass
- This is implemented via a filter based on character n-gram overlap of pairs of names

canonical name of i^{th} entity	$e_i \sim DP(H, \gamma)$
context vector of i^{th} entity	$\theta_i \sim Dir(\alpha)$
string of the j^{th} mention	$m_j \sim Transducer(e_{P(j)})$
context words of the j^{th} mention	$c_j \sim Mult(\theta_{P(j)})$



Gibbs sampling distribution

$$\begin{aligned} & p(m \mid e_i) \times p(c \mid \theta_i) \times p(e_i, \theta_i \mid H, \gamma, e_{-i}, \theta_{-i}) \\ & \quad \propto \exp(\vec{w} \cdot f(\text{“Thomas Jefferson”}, \text{“Mr. Jefferson”})) && \text{// Log-linear name transducer model} \\ & \quad \times DirMult(c \mid \alpha) && \text{// Multinomial context language model} \\ & \quad \times \frac{M_i + \alpha}{M - 1 + \alpha} && \text{// Dirichlet process prior} \end{aligned}$$

Data

- We use a 2008 dump of Wikipedia for evaluation
- Filter Wikipedia articles down to people listed on Freebase
- Mentions are the anchor texts of links that point to these entities' pages
- 1M entities and 22M mentions
- We are in the process of evaluating our system