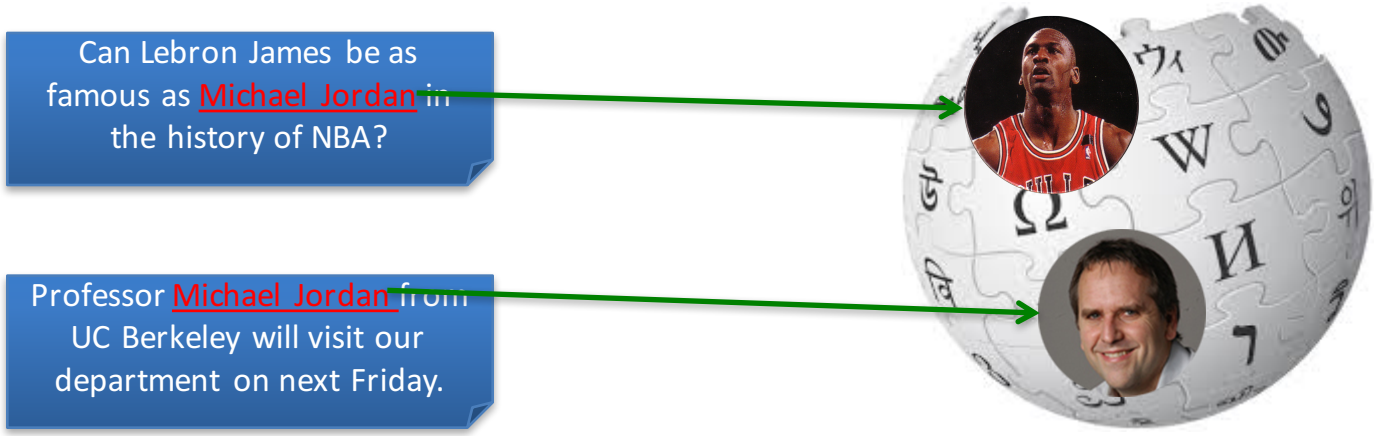
Entity Disambiguation with Linkless Knowledge Bases

Yang Li, Shulong Tan, Huan Sun, Jiawei Han, Dan Roth and Xifeng Yan

Dept. of Computer Science, UC Santa Barbara
Dept. of Computer Science, UIUC

Named Entity Disambiguation (NED)

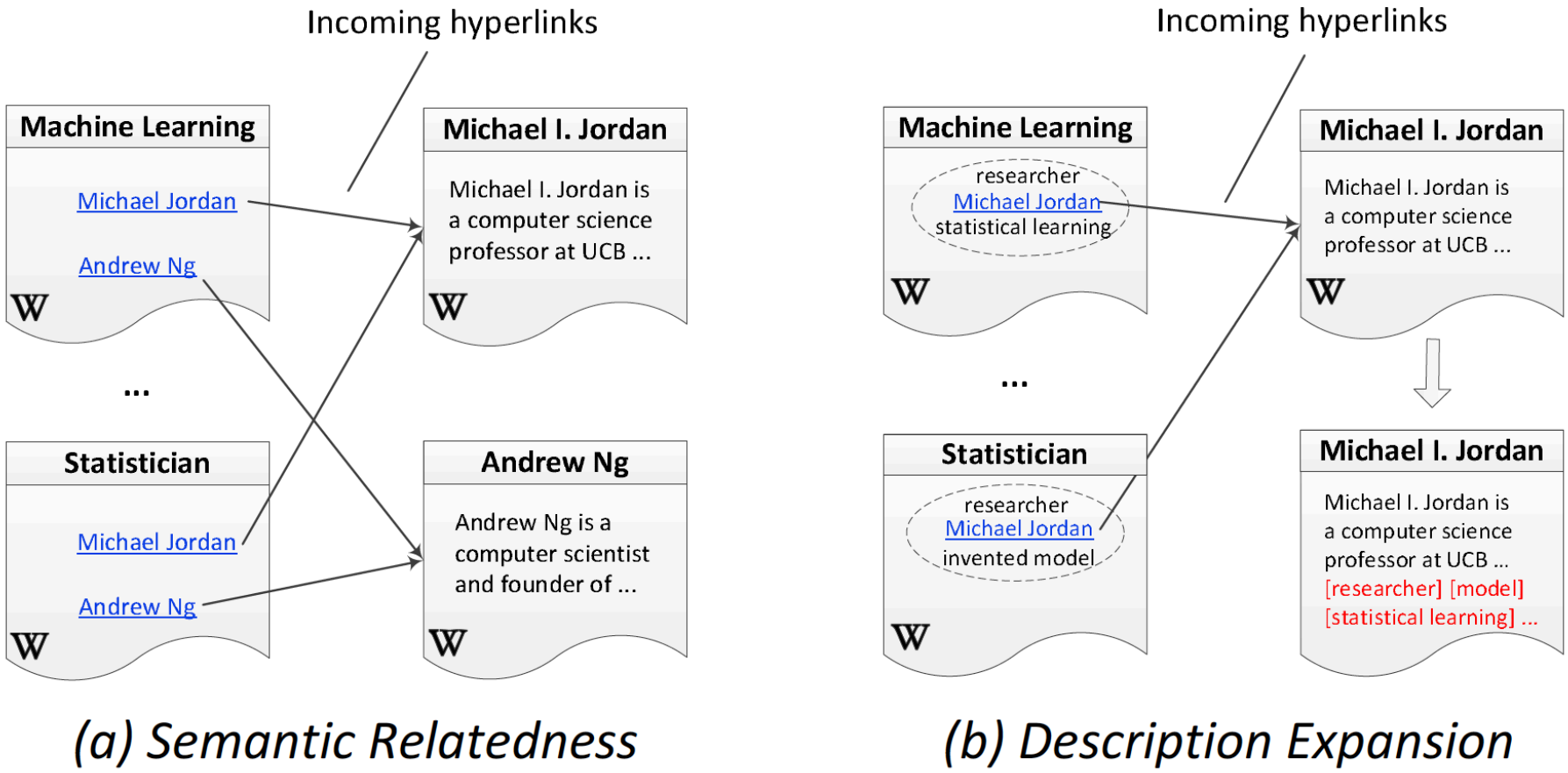
- Goal: map entity mentions in text to their corresponding entities in a reference Knowledge Base (e.g. Wikipedia).



- NED is critical for many text analysis/understanding tasks.
  - information network construction
  - tweets tagging
  - advertisements placement

NED - Existing solutions

- Heavily rely on the **cross-document hyperlinks in KB.**



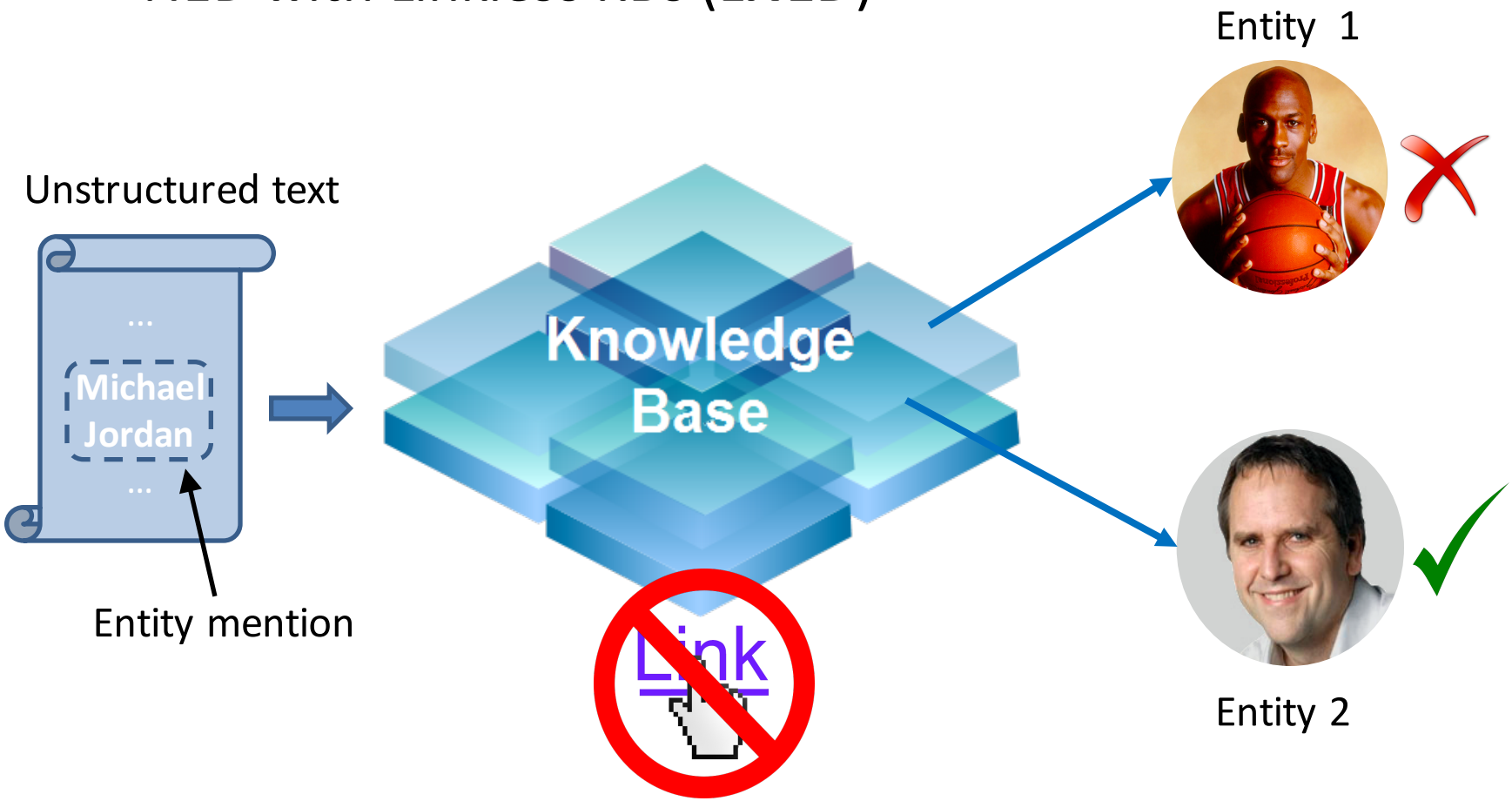
Motivation

- However…
  - Most closed domain KBs contain very few such links
    - Biomedicine
    - Enterprise
  - Manually adding such links into KB is very expensive
- So…

Is it possible to perform high-quality NED without using any cross-doc hyperlinks?

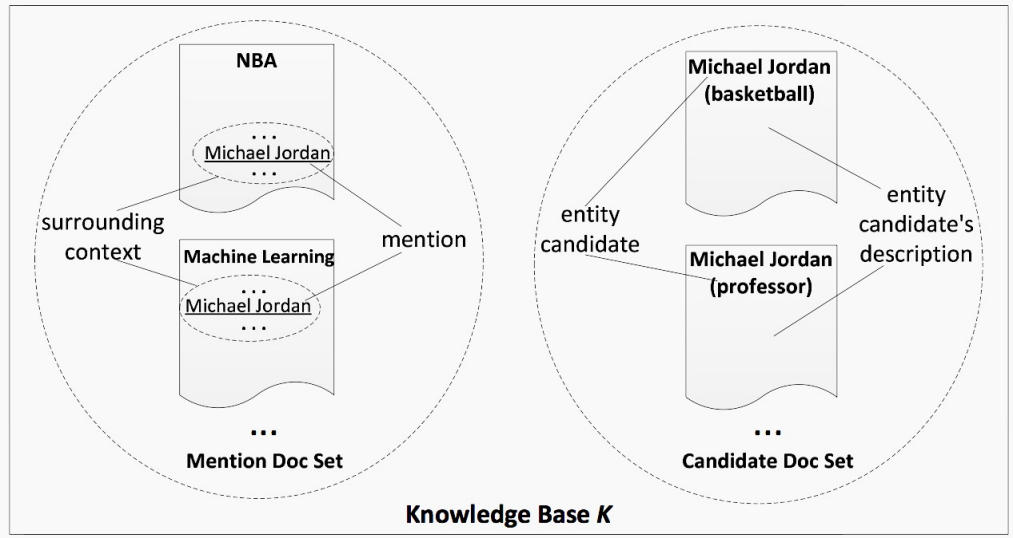
Objectives

- NED with Linkless KBs (**LNED**)



The Evidence Mining Approach

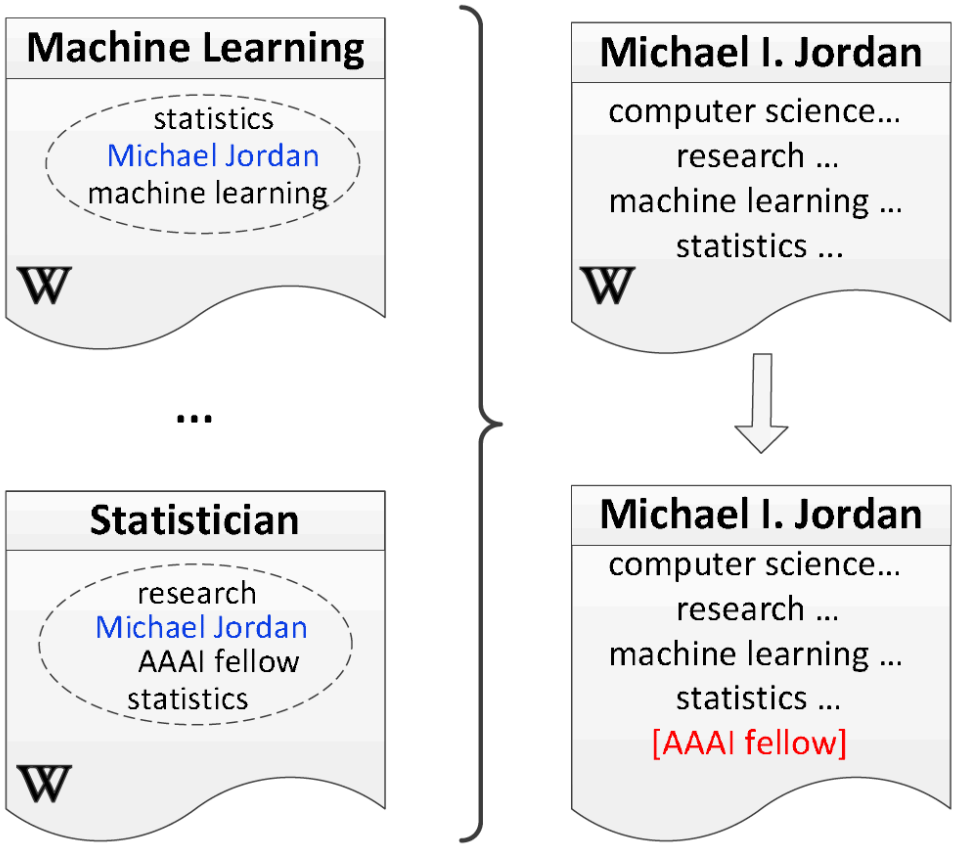
- Goal: bridge the information gap caused by missing links.
  - Input:
    - mention *m* and linkless reference KB *K*
    - *m*'s *candidate documents* and *mention documents*



  - Output:
    - A word distribution for each entity candidate (i.e., disambiguation evidences), with representative words higher probabilities

Disambiguation Evidences

- Mined evidences can expand the description of an entity.



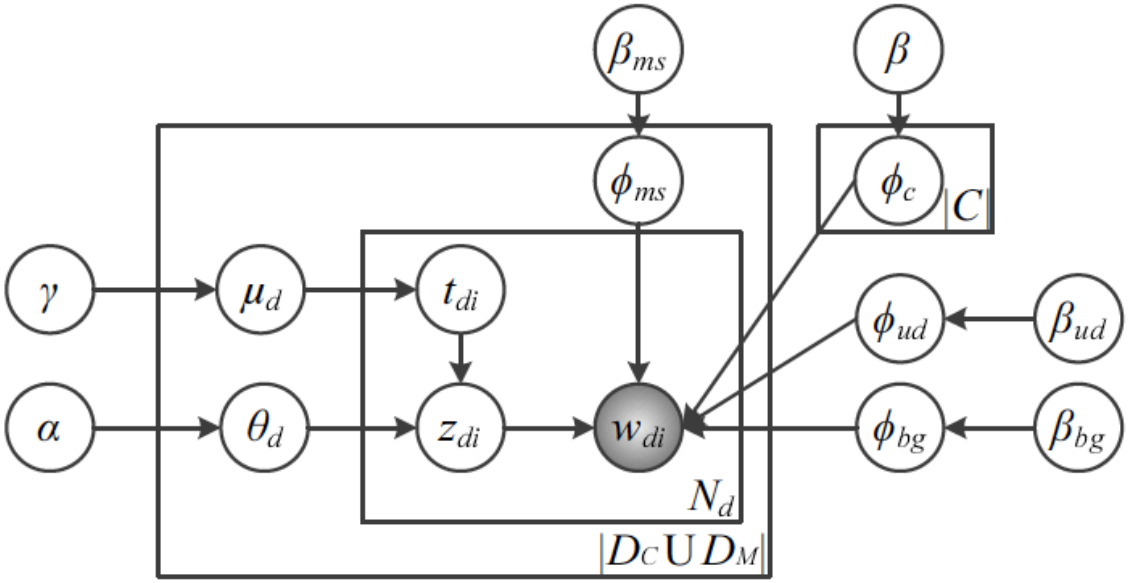
LNED via Evidence Mining

Algorithm 1 LNED via Evidence Mining

Input: Reference knowledge base K (with no links), named entity mention m, query q.

1: Generate candidates list C for mention m
2: Fetch candidate documents set D_C from K
3: Fetch m's mention documents set D_M from K
4: Mine evidences from $D_C \cup D_M$
5: Use mined evidences to rank candidate $c \in C$ for m in q
6: Return top-ranked candidate c_{top} as the genuine entity for m in q

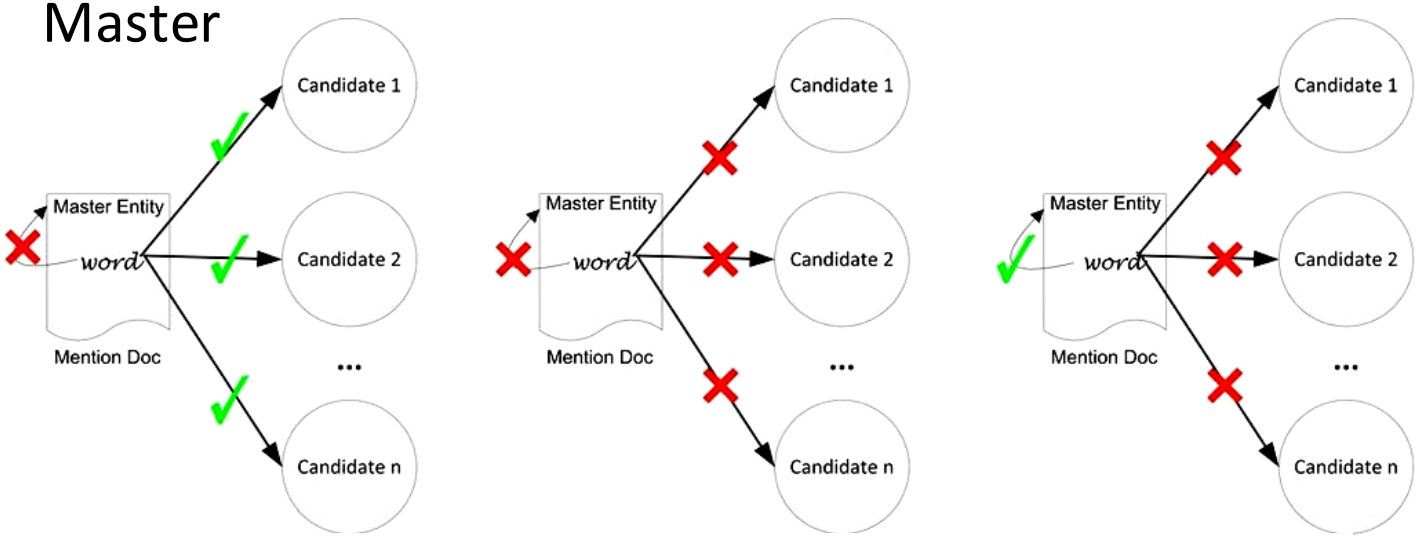
Evidence Mining Model



- A Generative Model
  - Given a target mention, D_C (candidate doc set), D_M (mention doc set)
    - model each of its entity candidates as a topic/label
    - introduce some special topics/labels to capture noisy/useless words
    - Generate the words in D_C & D_M based on such topics

Evidence Mining Model

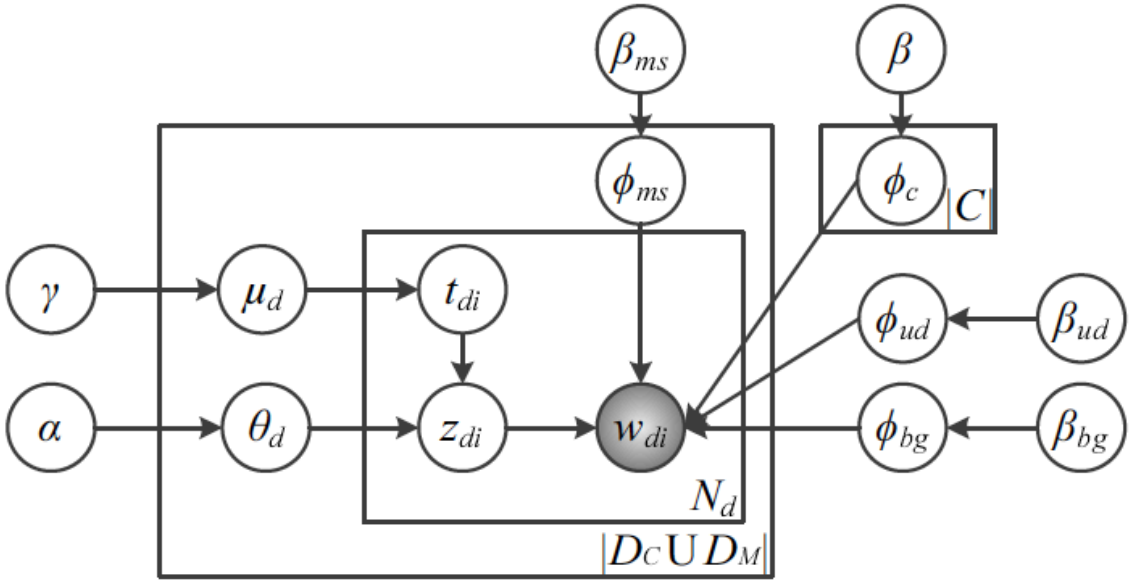
- Three special types of topics/labels:
  - Background
  - Undefined
  - Master



(a) Background (b) Undefined (c) Master

- For a mention with K referent entity candidates, the total number of topics/labels is K+2+| master entity set | or K+2+| mention document set|

Evidence Mining Model



$for\ w_{di} \in D_C$	$for\ w_{di} \in D_M$
z_{di} is either the corresponding candidate label, or "background"	For words surrounding mention (width-*W* window): z_{di} is either drawn from the referent entity candidates' labels plus "undefined", or "background", or "master" For other words: z_{di} is "master"

Model Inference

- Approximate Inference via Gibbs Sampling:
  - Blocked Gibbs Sampling
  - Sample $\{z_{di}, t_{di}\}$ together given all other variables

- Estimating Document-Label Association:

$$\theta_d^{(c)} = \frac{|w \in d, t_w = 1, z_w = c| + \alpha_c}{|w \in d, t_w = 1| + |C| \cdot \alpha + \alpha_{ud}}$$

- Estimating Label-Word Association:

$$\phi_c^{(v)} = \frac{|w = v, t_w = 1, z_w = c| + \beta}{|t_w = 1, z_w = c| + V \cdot \beta}$$

Ranking Referent Candidates

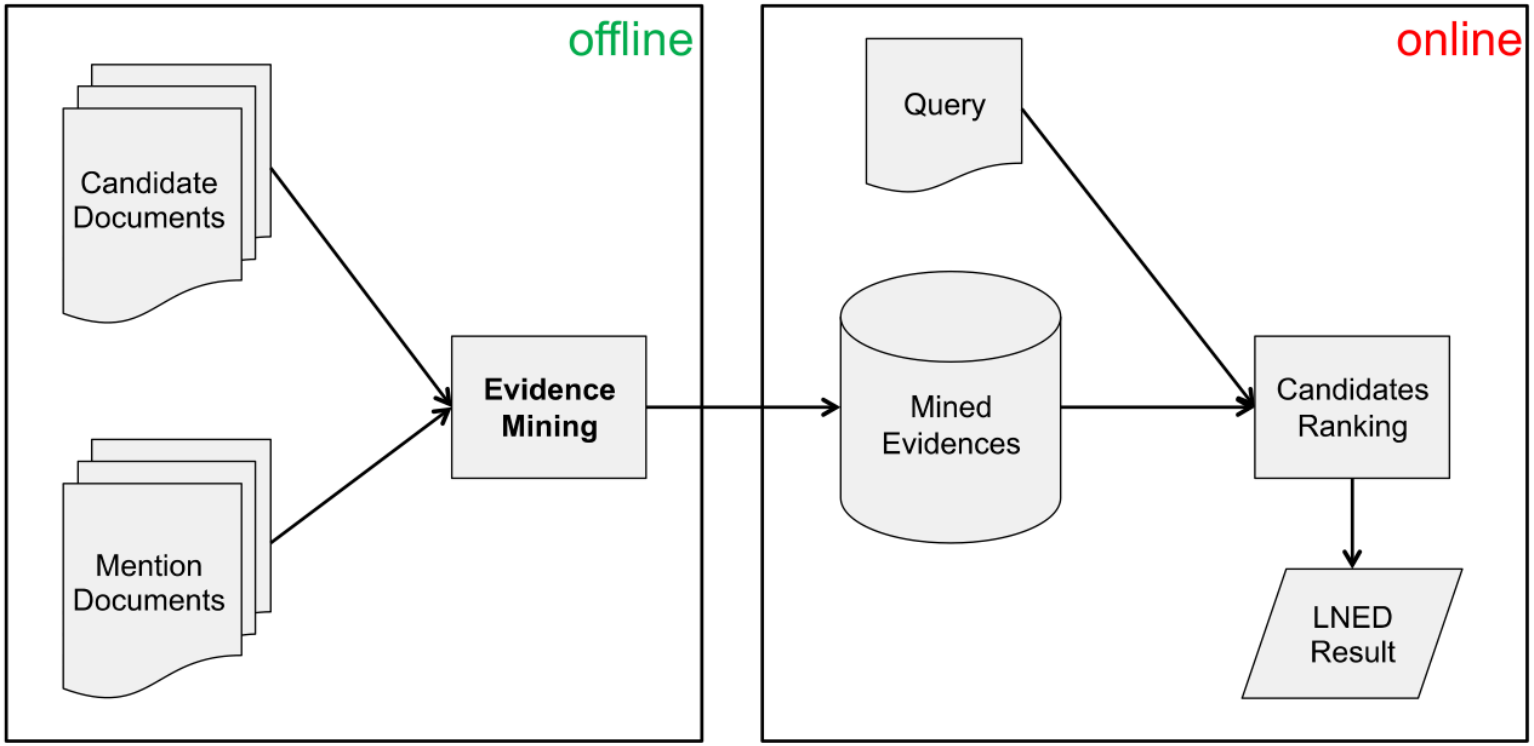
- Utilize the knowledge learned from the evidence mining model to rank referent entity candidates, and choose the top-ranked candidate as disambiguation result.

- Via Incremental Gibbs Sampling:
  - only sample the words in the query document
  - converge very fast

- Predict with Maximal Marginal Probability

$$LNED(d) = argmax_c \theta_d^{(c)}$$

LNED via Evidence Mining



Experiments Setup

- Datasets

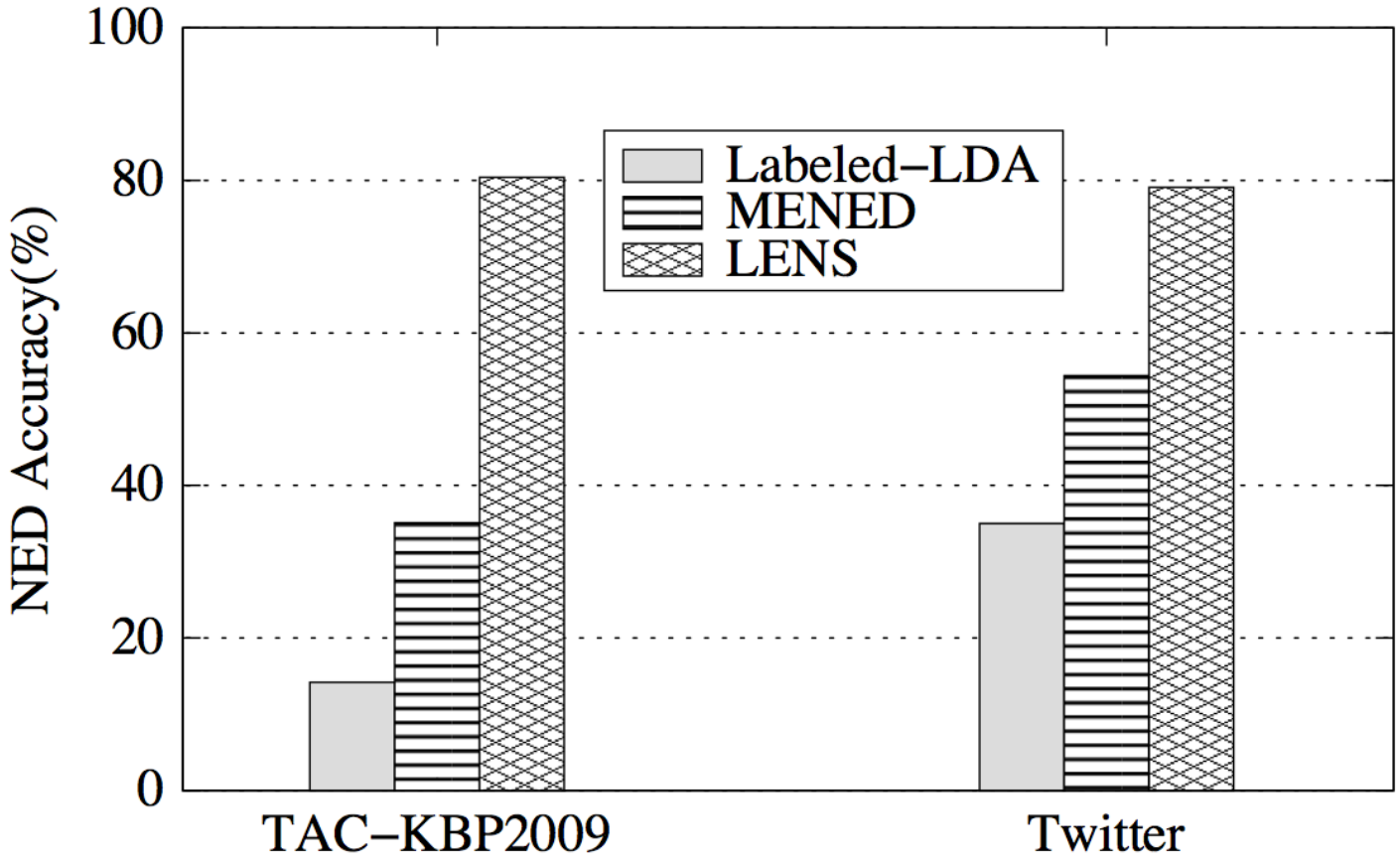
	TAC-KBP2009	Twitter
# of Queries	424	340
Avg Length of Queries (words)	53.15	16.46
Avg # of Candidates	~24	~19

- Reference Knowledge Base
  - Wikipedia (with all hyperlinks removed)
- Parameter Setting
  - tuned on a small test dataset
  - $\alpha = 0.01, \alpha_{df} = 0.1, \beta = 0.01, \beta_{df} = 0.1, \beta_{bg} = 0.1, \beta_{ms} = 0.01$
  - $\gamma_1 = 0.01, \gamma_2 = 1, \gamma_3 = 2$

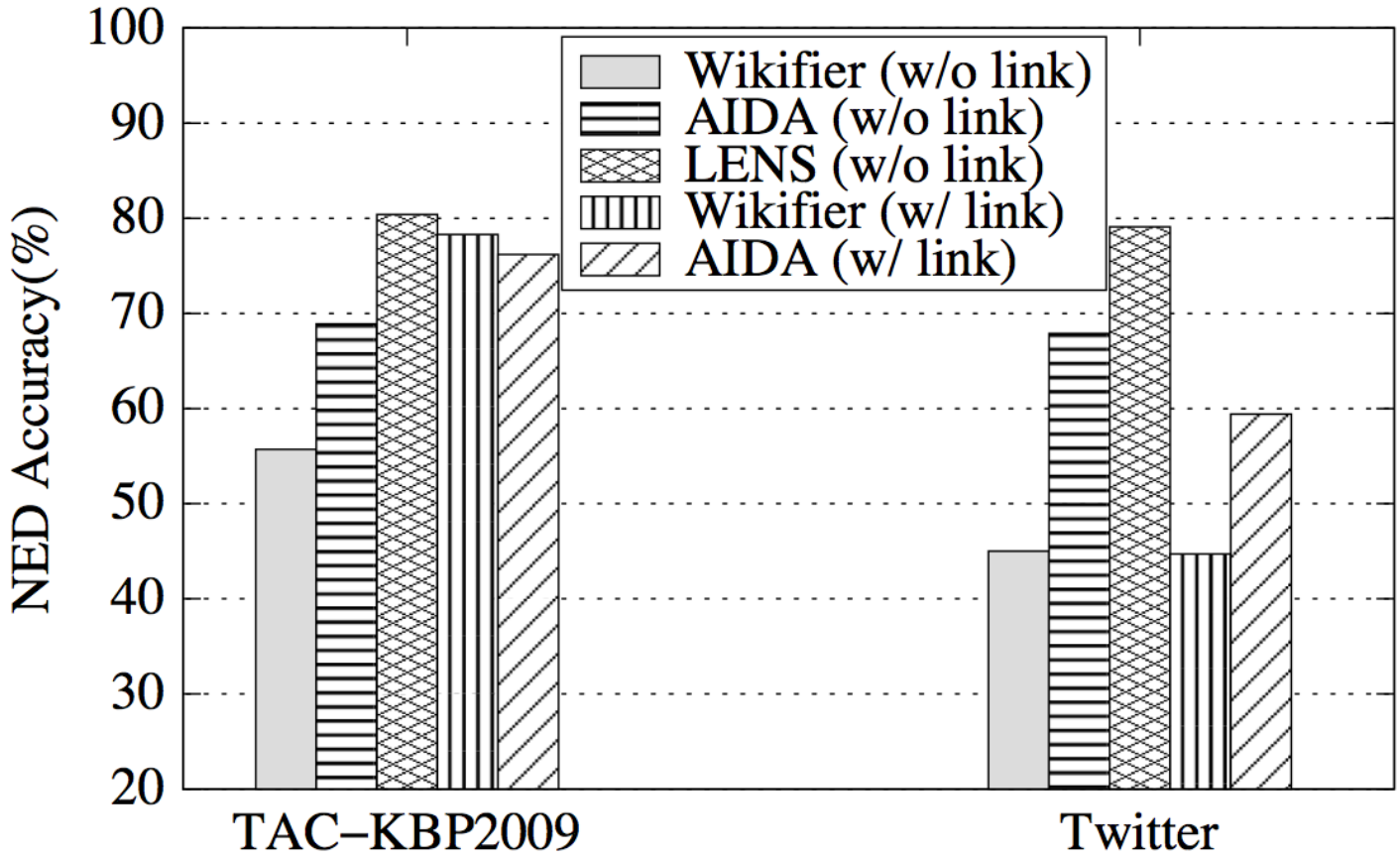
Experiments Setup

- Compared methods:
  - ***Labeled-LDA***: a model which learns label-word association from labeled documents and infers labels for unlabeled documents. [1]
  - ***MENED***: a model designed to mine additional evidences from external corpus to help NED. [2]
  - ***Wikifier***: a widely-used NED system using a machine learning based hybrid strategy to combine various kinds of features. [3]
  - ***AIDA***: a robust NED system making use of weighted mention-entity graph to find the best joint mention-entity mapping. [4]
  - ***LENS***: our method, we name it as **L**inking **E**vidences in **N**ot well linked **S**ources (LENS).

Effectiveness of Evidence Mining



End-to-end NED Accuracy



Conclusions

- Named Entity Disambiguation with Linkless Knowledge Bases (LNED)
  - LNED is a critical and challenging task, especially in domains of biomedicine, enterprise, etc.
  - Our evidence mining approach provides an effective way to tackle the LNED problem.

- Future work
  - Investigating possibility to test in closed domains
  - Automatically generating entity candidates without relying on any mention-entity mapping dictionaries.

References

[1] D. Ramage *et. al*, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora"

[2] Y. Li *et. al*, "Mining evidences for named entity disambiguation"

[3] L. Ratinov *et. al*, "Local and global algorithms for disambiguation to Wikipedia"

[4] J. Hoffart *et. al, "*Robust disambiguation of named entities in text"