Mining Evidences for Named Entity Disambiguation

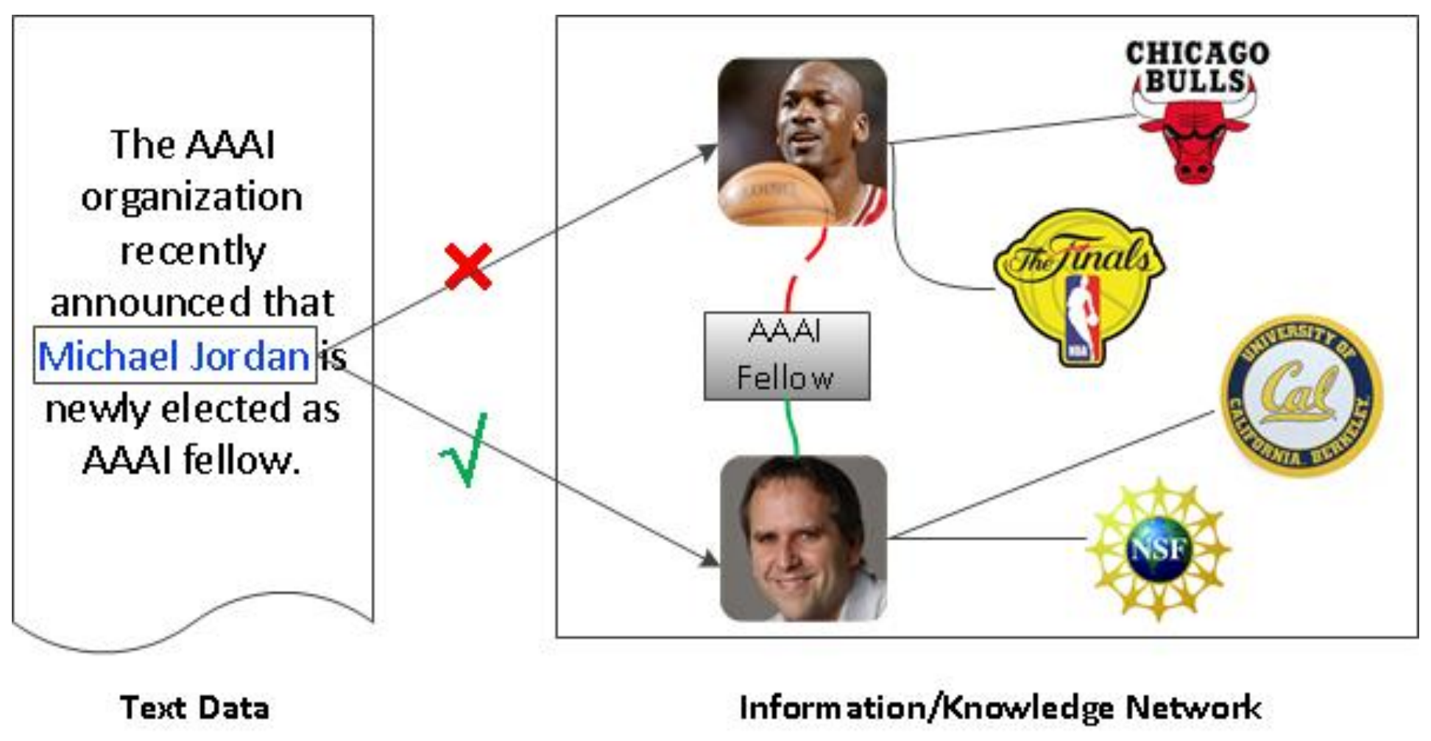
Yang Li[1], Chi Wang[2], Fangqiu Han[1], Jiawei Han[2], Dan Roth[2], Xifeng Yan[1]

[1] Department of Computer Science, UCSB, CA 93106, USA

[2] Department of Computer Science, UIUC, IL 61801, USA

Introduction

- Named Entity Disambiguation (**NED**) is an important component in constructing high-quality information network or knowledge graph from text.


The AAAI organization recently announced that Michael Jordan is newly elected as AAAI fellow.
CHICAGO BULLS
AAAI Fellow
Text Data
Information/Knowledge Network

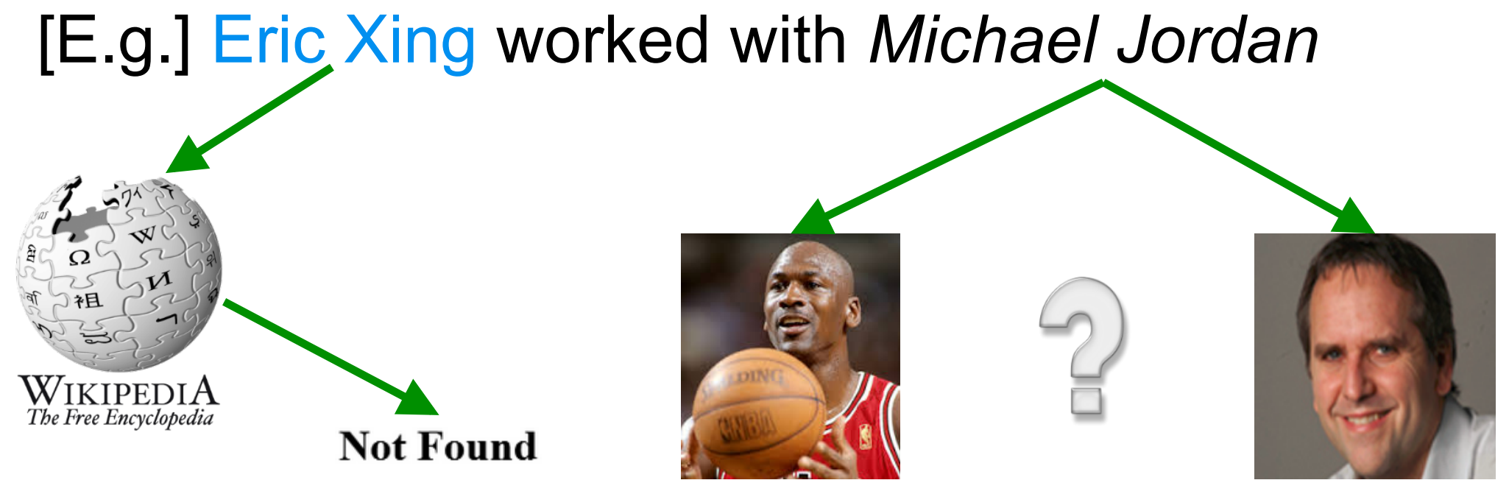
- Previous research on NED assumes that the reference knowledge base can provide enough explicit information to help disambiguate a mention to the right entity, which is not true in most cases, thus leading to poor performances on short queries with not well-known contexts.
- We introduce a novel task, **mining evidences for NED**, to collect additional evidences scattered in internal/external corpus to augment the knowledge bases and enhance their disambiguation power.
- We propose a generative model and an incremental algorithm to automatically mine useful evidences across documents. The mined evidences can help boost the disambiguation performance significantly.

Mining Evidences for NED

Aims at Solving:

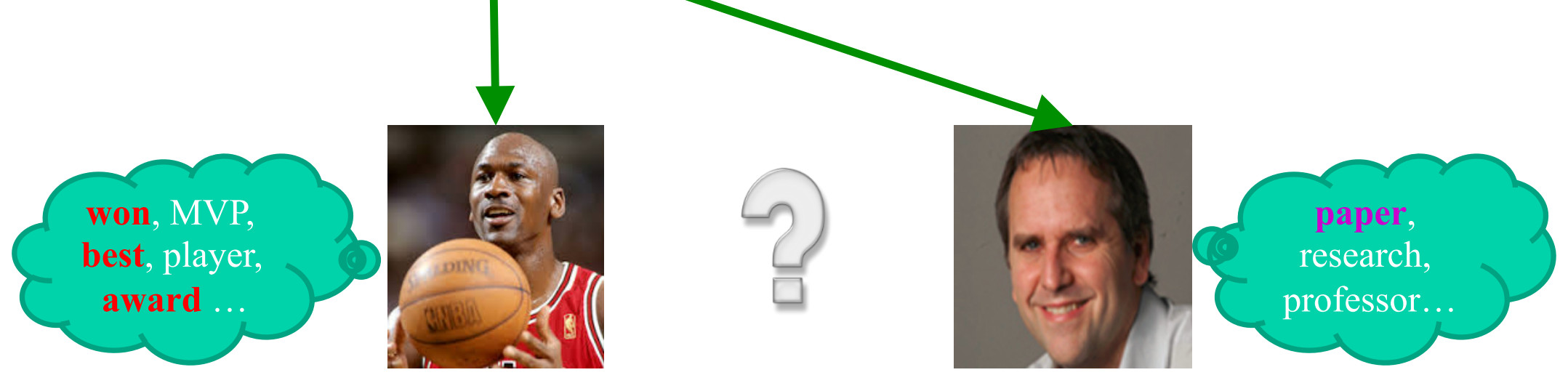
a) No evidence failure

[E.g.] Eric Xing worked with *Michael Jordan*



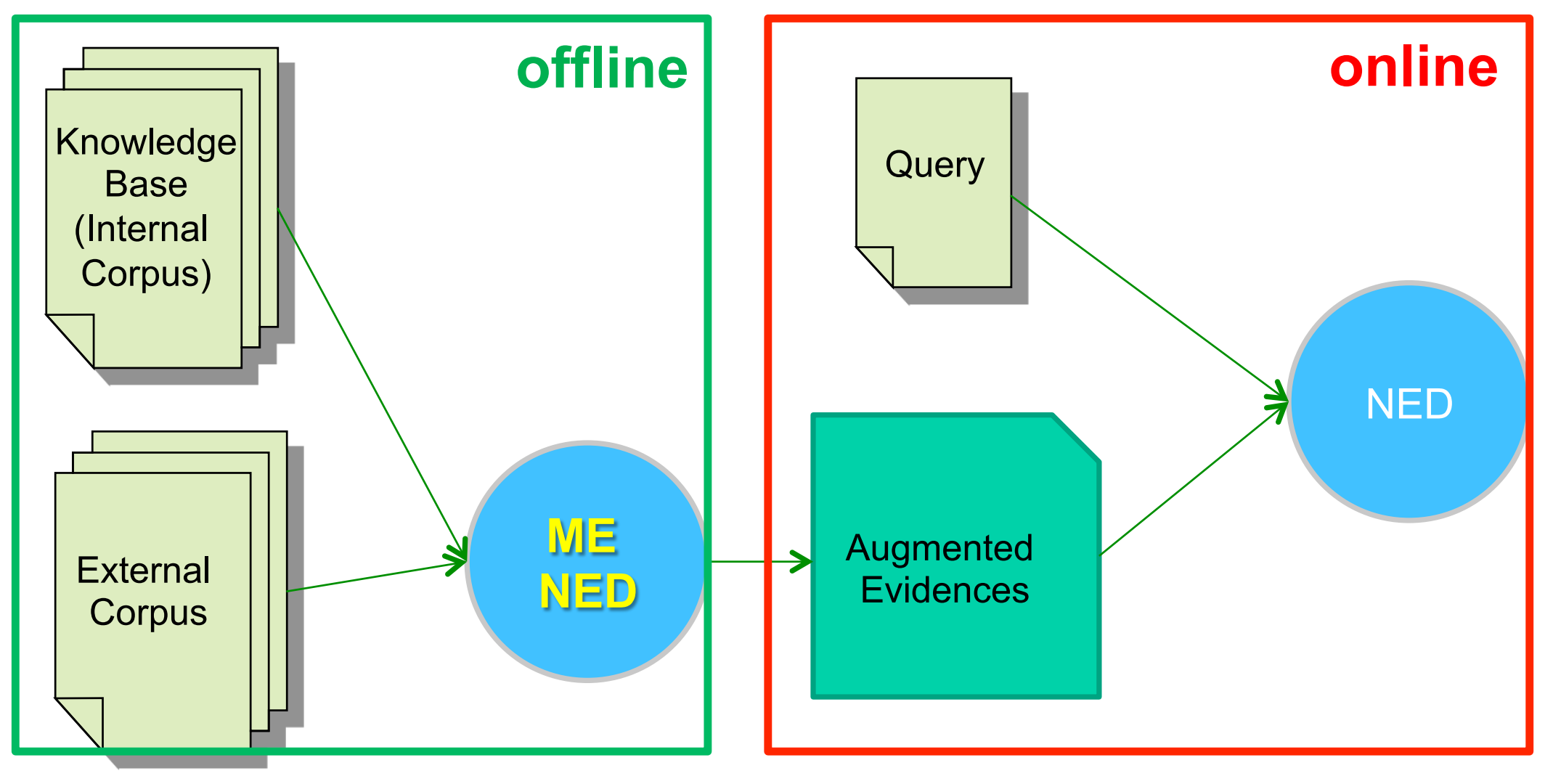
b) Insufficient evidence failure

[E.g.] *Michael Jordan* won the best paper award



Overview:

- Mining Evidences for NED (a.k.a **MENED**)
  - a) independent of query context
  - b) run offline as a preprocessing step.

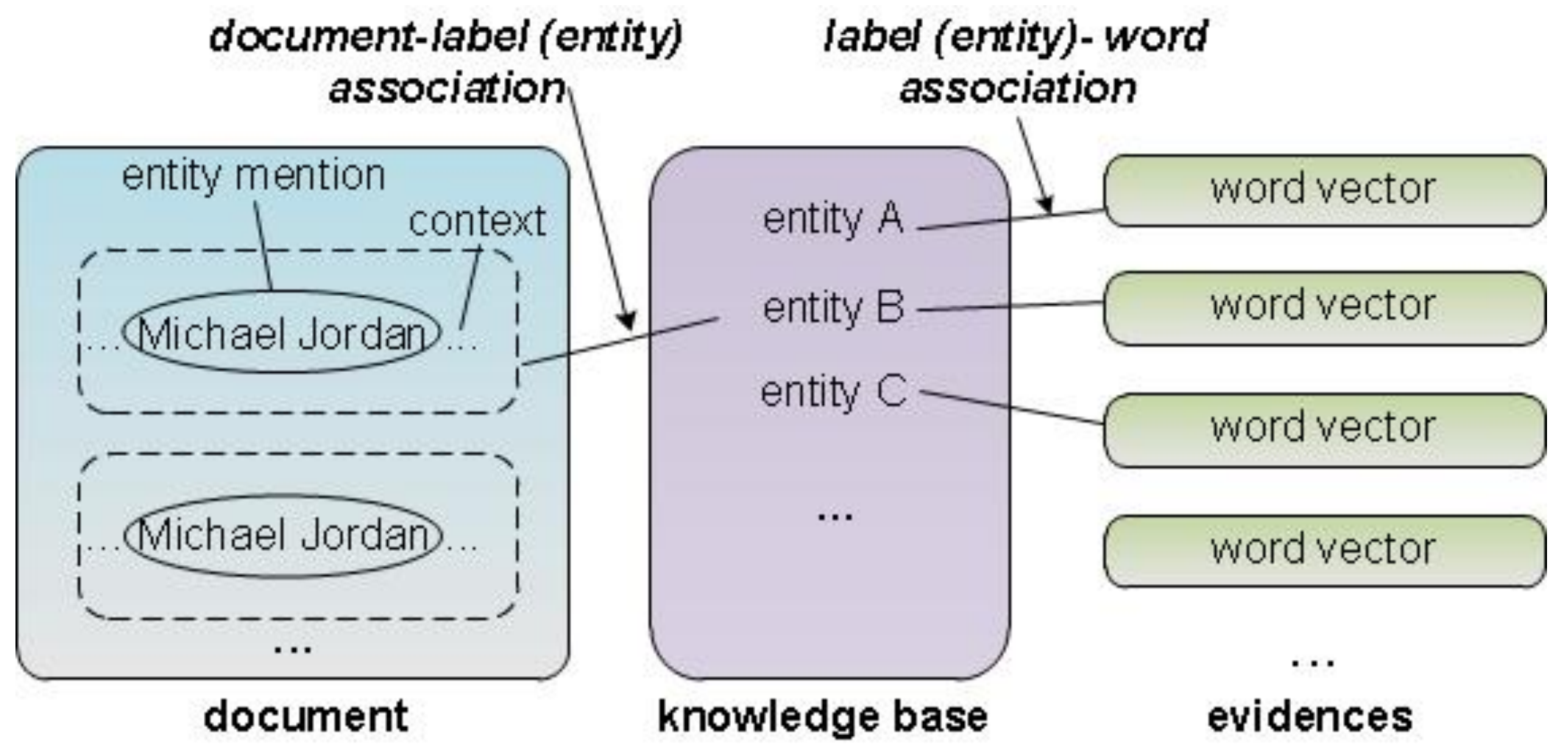


Method

Intuition

- Use labeled/linked docs in Wikipedia as initial evidences
- Search for a set of unlabeled docs $D_{external}$ from Google
- Jointly disambiguate and extract evidences from $D_{external}$

Model Basics

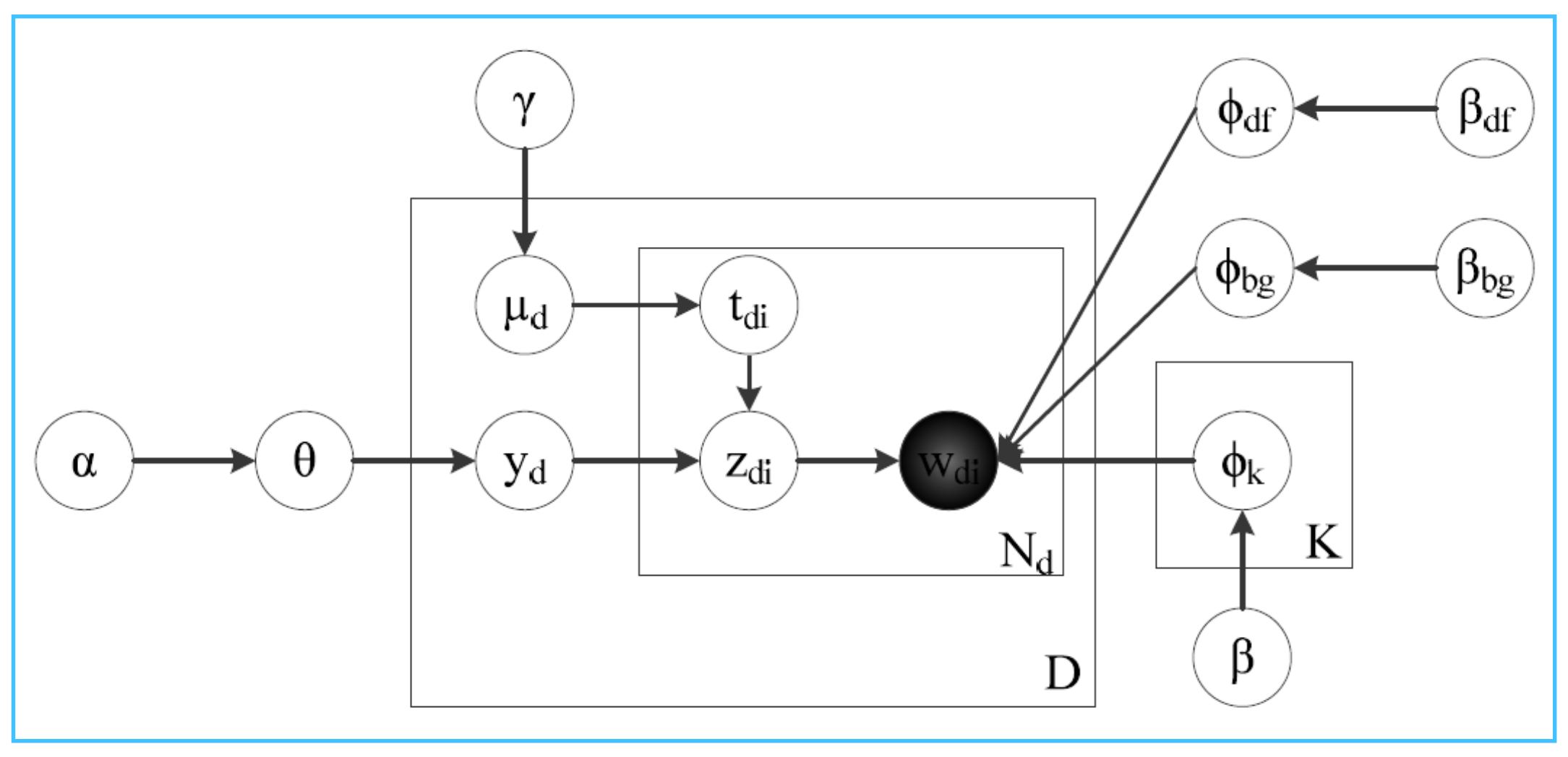


Two Special Topics/Labels

a) Background: capture words being general to topics

b) Default: capture words not belonging to any topics

A Generative Model

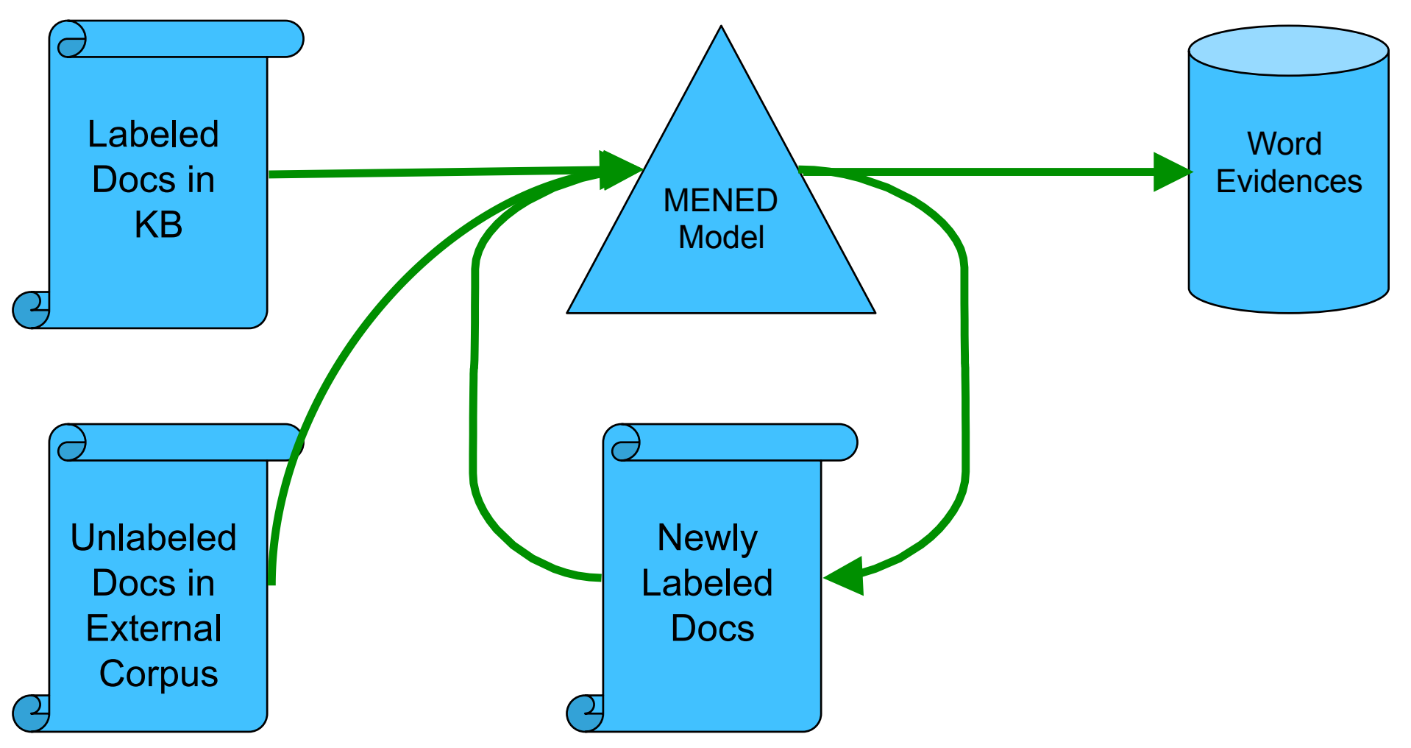


- Different Dirichlet priors for regular/bg/df topics
- Each document has only one topic/label
- Word label is restricted by document label

Model Inference

- Via Blocked Gibbs Sampling
- With Variational Approximation
- Estimating Document Label y_d & Word Label z_{di}

Incremental Evidence Mining



Experimental Study

Setup

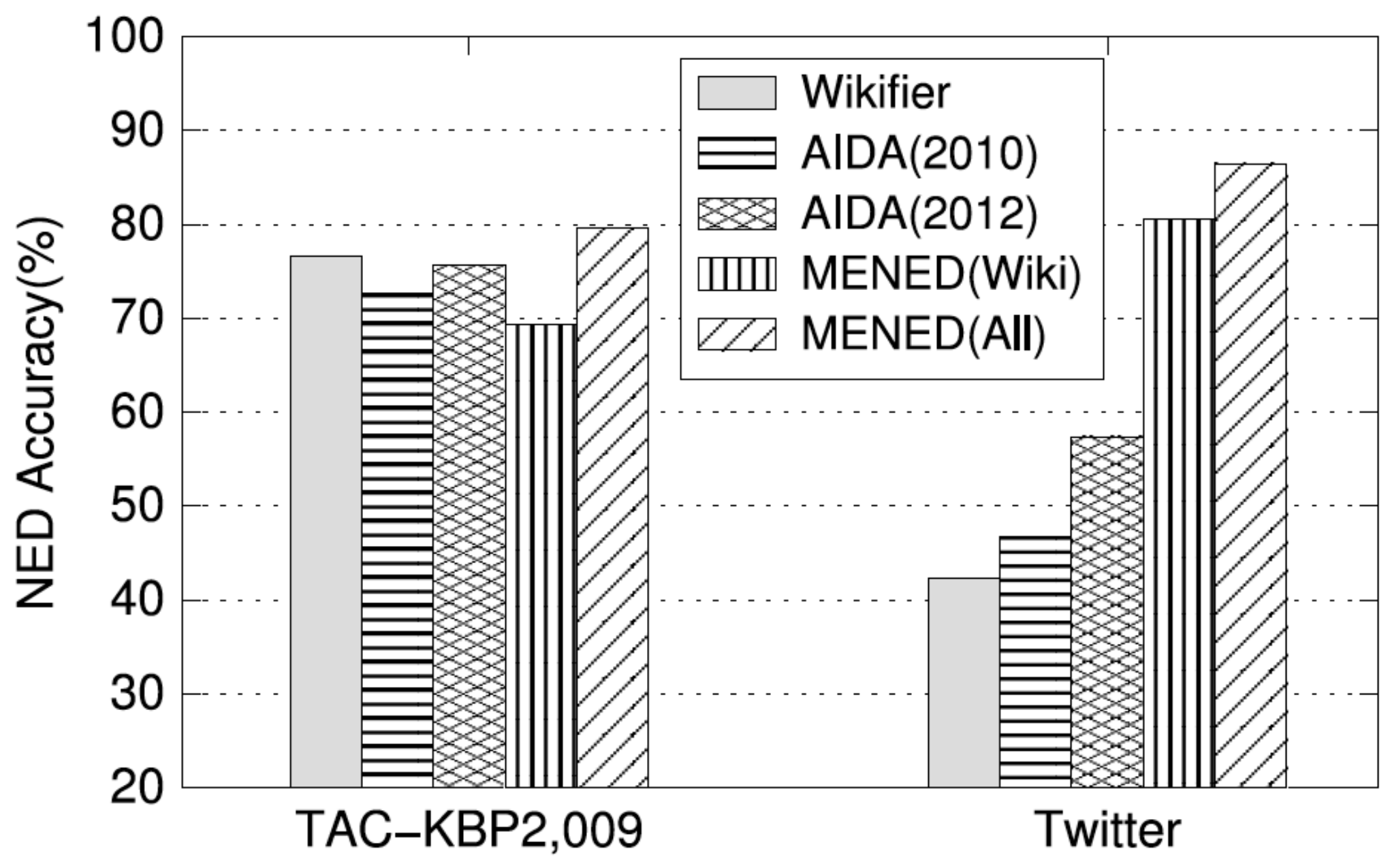
- *Datasets:*

Datasets	# Queries
TAC-KBP 2009	424
Twitter	340

- *Reference Knowledge Base:* Wikipedia
- *External Corpus:* documents indexed by Google
- *Baselines:* Wikifier [1] and AIDA [2]

Experimental Study (Cont'd)

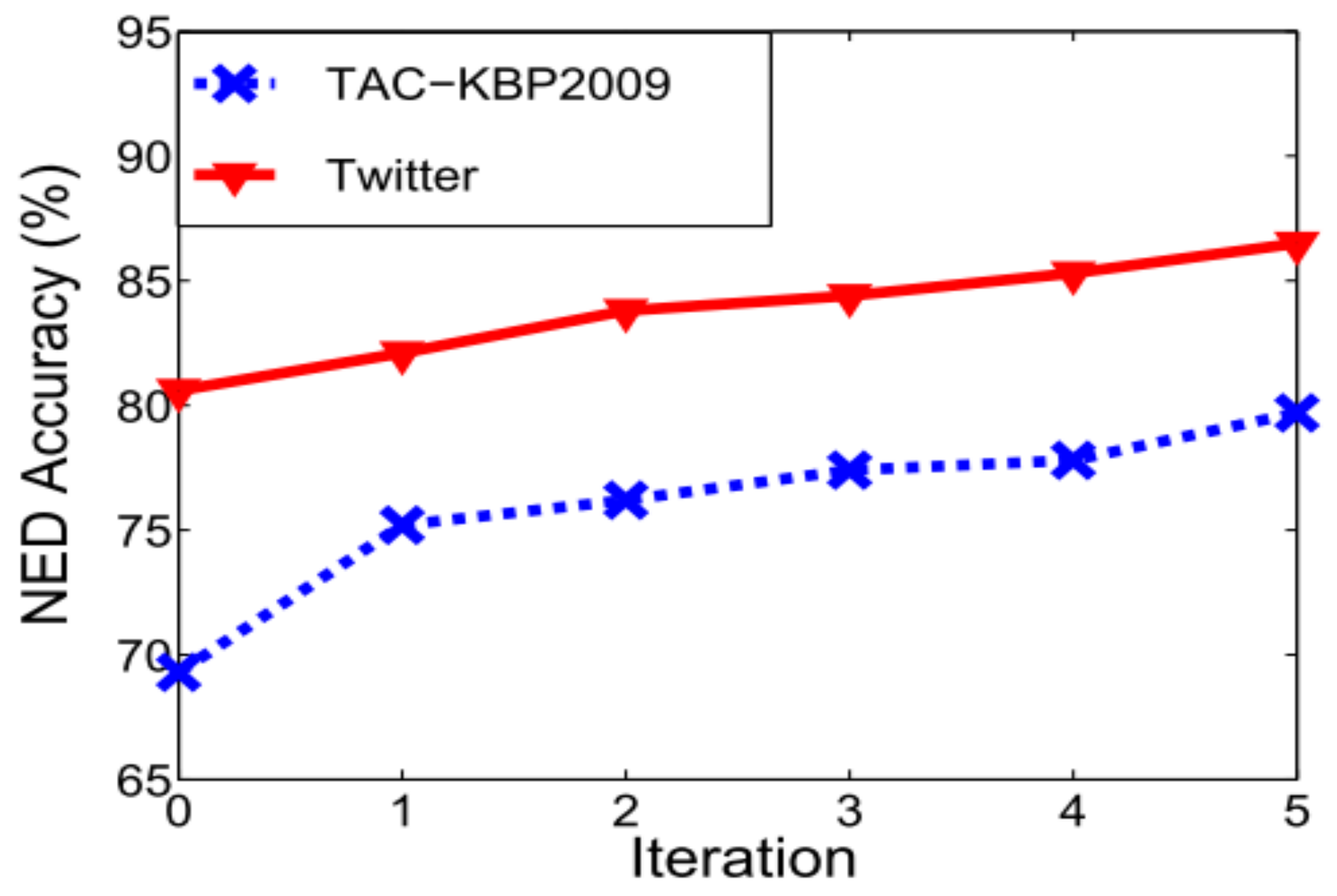
Effectiveness of Evidence Mining:



Sample Evidences Mined outside KB:

Entity	Mined Additional Evidences
Michael I. Jordan (Michael Jordan)	layers, nonparametric, nonlinear, pehong, chen, distinguished, david, heckerman, kearns, marina, meila, ...
Michael B. Jordan (Michael Jordan)	wood, oscar, role, peters, gilliard, detmer, larry, freamon, true-frost, pryzbylewski, octavia, spencer, troubled, ...
Owen Bieber (Bieber)	jobs, automobile, corporation, approved, presidential, lofton, support, vote, organizer, worley, conventions, ...
Aircraft Hotspur (Hotspur)	operating, ground, states, cargo, aviation, capacity, built, fighter, targets, spitfire, flight, eben, paratroops ...
David Y. Cameron (David Cameron)	engravers, technique, sculpture, printmaking, reproduced, scotch, lorne, muirhead, walton, french, nature, ...

Impact of Incremental Mining:



Conclusions

- Mining additional evidences to augment knowledge base is necessary for improving NED performance
- Our proposed generative model and incremental algorithm are effective in performing MENED
- This work yields a promising method to fill the information gap between the knowledge base and the NED query. As future work, we plan to extend our approach to mine other type of evidences such as phrases and concepts.

References

[1] Lev Ratinov *et al.* "Local and global algorithms for disambiguation to wikipedia". In *ACL* 2011.

[2] Johannes Hoffart *et al.* "Robust disambiguation of named entities in text. In *EMNLP* 2011.