Mining Evidences for Named Entity Disambiguation

Yang Li¹, Chi Wang², Fangqiu Han¹, Jiawei Han², Dan Roth², Xifeng Yan¹

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 networks or knowledge graphs. 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/unlabeled docs in Wikipedia as initial evidences
- Search for a set of unlabeled docs \( D_{\text{external}} \) from Google
- Jointly disambiguate and extract evidences from \( D_{\text{external}} \)

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 \( z_{dw} \)

Incremental Evidence Mining

Experimental Study (Cont’d)

Effectiveness of Evidence Mining:

- Sample Evidences Mined outside KB:
  - Michael Jordan (Michael Jordan)
  - wood, oscar, role, petes, gilliard, detmer, larry, freeman, true-frost, pezybylowsk, octavia, spencer, trebland, ...
  - jobs, automobile, corporation, approved, presidential, lexicon, support, vote, organizer, worley, conventions, ...
  - operating, ground, states, cargo, aviation, capacity, built, fighter, targets, spitfire, ...

- Impact of Incremental Mining:
  - TAC-KBP2009 vs. Twitter

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.