NERank: Bringing Order to Named Entities from Texts

Chengyu Wang[1], Rong Zhang[1], Xiaofeng He[1], Guomin Zhou[2], Aoying Zhou[1]

[1] Institute for Data Science and Engineering, East China Normal University

[2] Zhejiang Police College

Outline

- **Introduction**
- Problem Statement
- Proposed Approach
- Experiments
- Conclusion

Entity Ranking

- Ranking entities from texts
  - Input: a text collection
  - Output: a ranked order of named entities
- Why entity ranking?
  - **Entity-oriented Web search**
    - given a query, retrieve a list of entities from relevant documents
  - **Web semantification**
    - add semantic tags to Web documents
  - **Knowledge base population**
    - extract and rank entities and then link them to knowledge bases

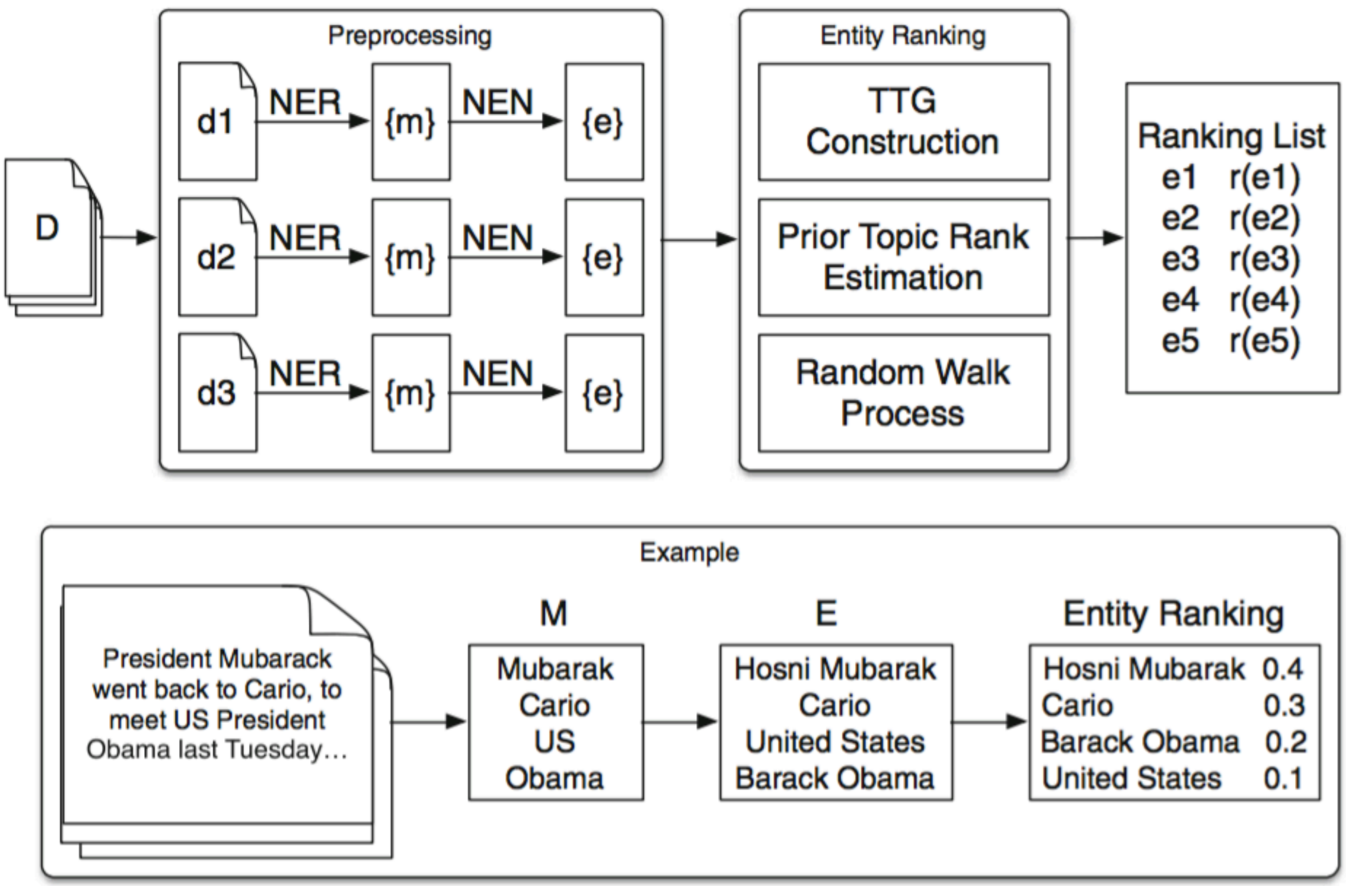
Outline

- Introduction
- **Problem Statement**
- Proposed Approach
- Experiments
- Conclusion

Problem Statement

- Given a document collection D and a normalized named entity collection E detected from D, the goal is to give each entity $e \in E$ a rank $r(e)$ to denote the relative importance such that
  - $0 \leq r(e) \leq 1$
  - $\sum_{e \in E} r(e) = 1$

General Framework



Outline

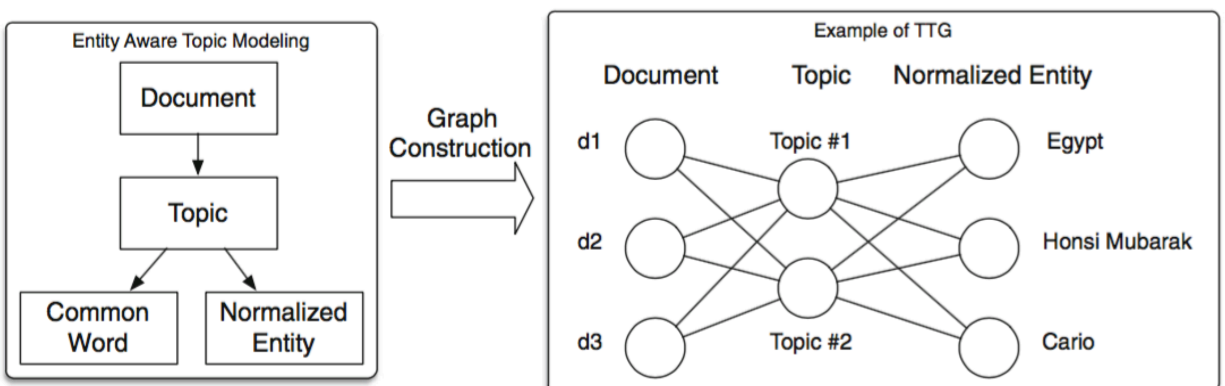
- Introduction
- Problem Statement
- **Proposed Approach**
- Experiments
- Conclusion

Topical Tripartite Graph Modeling

- Topics in Egypt Revolution

Topic	Top normalized entities	Top common words	Description
#1	Egypt, Hosni Mubarak	political, military, revolution	Start of the revolution
#2	Mohamed Morsi, Egypt	President, constitution, vote	Presidential election
#3	Egypt, Israel, Iran	government, foreign, peace	Foreign countries' reaction
#4	Egypt, Cairo	economic, government, billion	Revolution's effect on economy
#5	Egypt	tourism, tourist, travel, sea	Revolution's effect on tourism

- TTG construction



Prior Topic Rank Estimation

Three Quality Metrics

- Probabilities derived from TTG modeling
  - $\theta_{i,j}$: probability of topic t_j in document d_i
  - $\hat{\varphi}_{i,j}$: probability of normalized entity e_j in topic t_i
- Quality metrics
  - Prior probability

$$pr(t_i) = \frac{1}{|D|} \sum_{j=1}^{|D|} \theta_{i,j}$$

  - Entity richness

$$er(t_i) = \frac{1}{Z_{er}} \sum_{j=1}^{|E|} \hat{\varphi}_{i,j}$$

  - Topic specificity

$$ts(t_i) = \begin{cases} 0, & (pr(t_i) < \varepsilon) \\ \frac{1}{Z_{ts}} \sum_{j=1}^{|D|} \theta_{i,j} \log_2 \theta_{i,j} & (pr(t_i) \geq \varepsilon) \end{cases}$$

Topic	Prior probability	Entity richness	Topic specificity
#1	0.184	0.159	0.146
#2	0.264	0.181	0.254
#3	0.110	0.116	0.074
#4	0.053	0.085	0.023
#5	0.017	0.039	0.007

Prior Topic Rank Estimation

Ranking Function

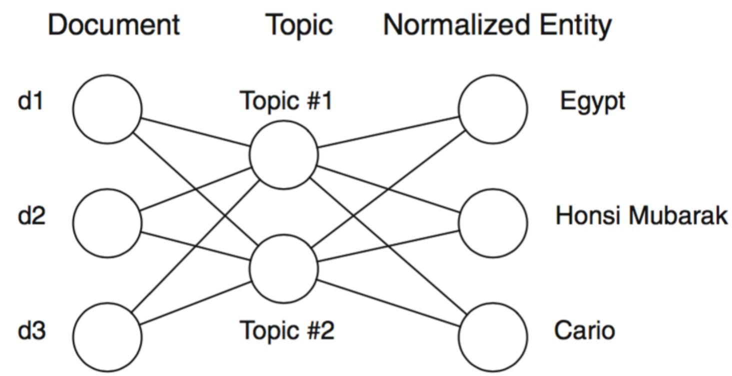
- Linear ranking function

$$r_0(t_i) = W^T \cdot F(t_i)$$

  - $F(t_i) =< pr(t_i), er(t_i), ts(t_i) >$
  - $\sum_i w_i = 1$

- Parameter learning
  - For two topics t_i and t_j, if t_i is a more important topic than t_j, we have $r_0(t_i) > r_0(t_j)$
  - Optimization objective: $\|W\|_2^2 + C \cdot \sum_{i,j} \xi_{i,j}$
  - Constraints: $W^T \cdot F(t_i) - W^T \cdot F(t_j) \geq 1 - \xi_{i,j}$
  - Train a linear SVM classifier to learn the weights

Meta-Path Constrained Random Walk Algorithm



- Initialization
  - $r(t_i) = r_0(t_i)$
- Probability propagation
  - Following TDT (Topic-Doc-Topic) meta path (with prob. $\alpha > 0$)

$$t_i \xrightarrow{\frac{\theta_{i,j}}{\sum_{d_k \in D} \theta_{k,j}}} d_j \xrightarrow{\theta_{j,k}} t_k$$

  - Following TET (Topic-Entity-Topic) meta path (with prob. $\beta > 0$)

$$t_i \xrightarrow{\frac{\widehat{\varphi}_{i,j}}{\sum_{e_k \in E} \widehat{\varphi}_{i,k}}} e_j \xrightarrow{\frac{\widehat{\varphi}_{k,j}}{\sum_{t_m \in T} \widehat{\varphi}_{m,j}}} t_k$$

  - Random jump (with prob. $1 - \alpha - \beta > 0$)

Proof of Convergence (1)

- Update rule of NERank

$$T_n = \alpha \cdot \Theta_R^T\Theta \cdot T_{n-1} + \beta \cdot \widehat{\Phi}_C\widehat{\Phi}_R^T \cdot T_{n-1} + (1-\alpha-\beta)T_0$$

- Non-recursive form of NERank

$$T_n = M^nT_0 + (1-\alpha-\beta)\sum_{i=0}^{n-1} M^iT_0$$

  – where $M = \alpha \cdot \Theta_R^T\Theta + \beta \cdot \widehat{\Phi}_C\widehat{\Phi}_R^T$

- Matrix limit of T_n
  – $\lim_{n\to\infty} T_n = \lim_{n\to\infty} M^nT_0 + (1-\alpha-\beta)\lim_{n\to\infty} \sum_{i=0}^{n-1} M^iT_0$
  – $\lim_{n\to\infty} M^nT_0 = 0$ (because $\Theta_R^T\Theta$ and $\widehat{\Phi}_C\widehat{\Phi}_R^T$ are transition matrices with $0 < \alpha + \beta < 1$)
  – $\lim_{n\to\infty} \sum_{i=0}^{n-1} M^iT_0 = (I-M)^{-1}T_0$

Proof of Convergence (2)

- Matrix limit of T_n

$$\lim_{n\to\infty} T_n = (1-\alpha-\beta)(I-M)^{-1}T_0$$

- Close form of T_n

$$T^* = (1-\alpha-\beta)(I-\alpha\cdot\Theta_R^T\Theta+\beta\cdot\hat{\Phi}_C\hat{\Phi}_R^T)^{-1}T_0$$

- Close form of E_n

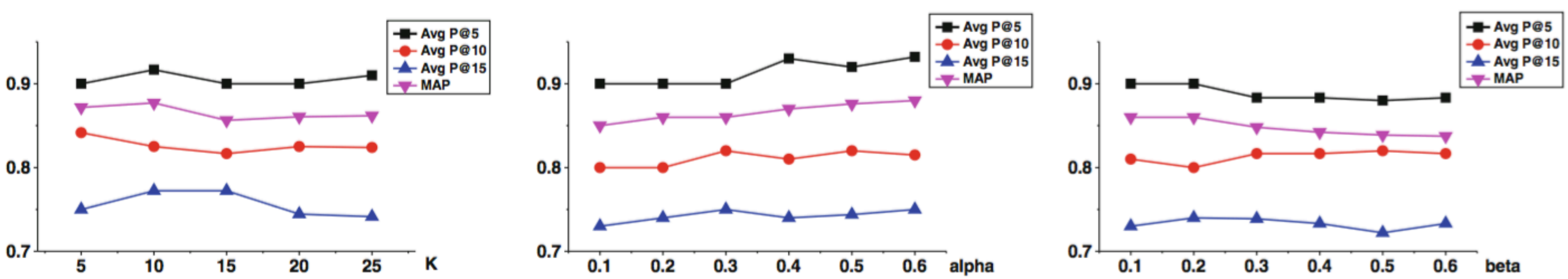
$$E^* = (1-\alpha-\beta)\widehat{\Phi}_R^T(I-\alpha\cdot\Theta_R^T\Theta+\beta\cdot\widehat{\Phi}_C\widehat{\Phi}_R^T)^{-1}T_0$$

Outline

- Introduction
- Problem Statement
- Proposed Approach
- **Experiments**
- Conclusion

Experiments (1)

- Datasets
  - 50 newswire collections from TimelineData and CrisisData, each related to an international event
  - Example events: Egypt Revolution, Iraq War, BP Oil Spill, etc.
- Hyper-parameter settings



(a) Varying $|T|$ (b) Varying α (c) Varying β

Experiments (2)

- Comparative study
  - Baselines: TF-IDF, TextRank, LexRank and Kim et al.
  - Variants of our approaches: $\text{NERank}_{\text{Uni}}$ and $\text{NERank}_{\alpha=0}$

Method	Average Precision@5	Average Precision@10	Average Precision@15	MAP
TF-IDF	0.85*	0.79*	0.73*	0.81*
TextRank	0.87*	0.83	0.73*	0.83*
LexRank	0.85*	0.8*	0.72*	0.8*
Kim et al.	0.87*	0.81*	0.76*	0.84*
NERank_{Uni}	0.80*	0.75*	0.71*	0.78*
$\text{NERank}_{\alpha=0}$	0.72*	0.61*	0.51*	0.62*
NERank	**0.92**	**0.87**	**0.79**	**0.89**

Experiments (3)

- Case studies

Entity	Egypt Revolution	Libya War	BP Oil Spill
1	Egypt	Libya	BP
2	Mohamed Morsi	Muammar Gaddafi	Gulf of Mexico
3	Hosni Mubarak	Tripoli	Barack Obama
4	Cario	NATO	Louisiana
5	Muslim Brotherhood	Benghazi	Coast Guard
6	Tahrir Square	Barack Obama	United States
7	Israel	Misrata	Tony Hayward
8	Middle East	United States	Deepwater Horizon
9	United States	National Transitional Council	Florida
10	Tunisia	Syria	Transocean

Outline

- Introduction
- Problem Statement
- Proposed Approach
- Experiments
- **Conclusion**

Conclusion

- NERank
  - Effective to rank named entities in documents with little human intervention
- Future work
  - A general framework for entity ranking from different types of texts (i.e., documents, tweets, etc.)
  - A complete benchmark for evaluating entity ranking

Thanks!

Questions & Answers