NERank: Ranking Named Entities in Document Collections

Chengyu Wang, Rong Zhang, Xiaofeng He, Aoying Zhou

Institute for Data Science and Engineering
East China Normal University
Shanghai, 200062, China

Introduction

- Named entity ranking is necessary to bring semantics to plain documents.
- Rank order of named entities should be determined by the relative importance considering the document collection.
- *NERank* is the first attempt to tackle the problem of named entity ranking directly from documents.

NERank Workflow

- Tripartite Graph Construction
  - The part aims to model the semantic relations between entities and documents indirectly by topic modeling. A weighted, tripartite graph is employed to represent $< document, topic, entity >$ relations.
- Prior Topic Rank Estimation
  - The part is responsible for assigning prior ranks to each topic in the tripartite graph.
- Random Walk Process
  - This part is designed to propagate prior topic ranks to documents and entities through a random walk process on the tripartite graph.

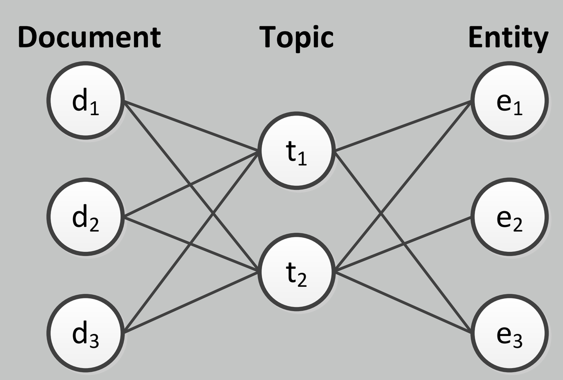

Figure 1: Tripartite Graph

Tripartite Graph Construction

- Named Entity Recognition and Normalization
  - Given a document collection D, perform NER and NEN to generate the entity set M and map each $m \in M$ to the normalized form $e \in E$.
- Entity-Aware Topic Modeling
  - Model a document $d \in D$ as the union set of common words and normalized named entities in E.
  - Estimate document-topic distribution $\mathbf{\Theta}$ and topic-word distribution $\mathbf{\Phi}$ by Gibbs sampling in LDA.
- Graph Construction
  - Nodes: documents D, topics T and entities E.
  - Edges: assign weights of $< document, topic >$ and $< topic, entity >$ edges by respective document-topic and topic-word probabilities.

Prior Topic Rank Estimation

- Estimate the prior rank for topic $t_i \in T$: $r_0(t_i)$.
- Three quality metrics:
  - Prior probability: the probability that topic t_i is discussed in D

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

where $\theta_{j,i}$ is the probability of topic t_i given document d_j.

  - Entity richness: the proportion of entities in words related to topic t_i

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

where $\phi_{i,j}$ is the probability of entity e_j given topic t_i, and Z_{er} is a normalization constant.

  - Topic specificity: whether the topic is specific about certain aspects or only provides background information

$$ts(t_i) = \frac{1}{Z_{ts}}\sum_{j=1}^{|D|}\theta_{j,i}\log_2\theta_{j,i}$$

where Z_{ts} is a normalization constant.

- Ranking topics by linear combination of quality metrics:

$$r_0(t_i) = \frac{1}{Z}(w_1 \cdot pr(t_i) + w_2 \cdot er(t_i) + w_3 \cdot ts(t_i))$$

where $Z = \sum_{t' \in T} r_0(t')$ is a normalization factor. $\forall i, w_i > 0$ and $\sum_i w_i = 1$. Weights are learned using a max-margin technique (a linear-SVM based supervised learning method).

Random Walk Process

- Select a topic $t_i \in T$ with probability $r_0(t_i)$ as the starting point.
- Make one of the following three transfers iteratively until the system reaches equilibrium (α and β are parameters where $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$):
  - With probability α, the random surfer walks through the path $t_i \to d_j \to t_k$. $d_j \in D$ is selected with probability $\frac{\theta_{j,i}}{\sum_{d_k \in D}\theta_{k,i}}$. Next, $t_k \in T$ is selected with probability $\theta_{j,k}$.
  - With probability β, the random surfer walks through the path $t_i \to e_j \to t_k$. $e_j \in E$ is selected with probability $\frac{\phi_{i,j}}{\sum_{e_k \in E}\phi_{i,k}}$. Next, $t_k \in T$ is selected with probability $\frac{\phi_{k,j}}{\sum_{t_m \in T}\phi_{m,j}}$.
  - With probability $1 - \alpha - \beta$, the random surfer jumps to a topic node t_j. t_j is selected with probability $r_0(t_j)$.
- Compute the rank of entity e_i:

$$r(e_i) = \frac{s(e_i)}{\sum_{e_j \in E} s(e_j)}$$

where $s(e_i)$ is the number of visits to e_i by random surfers.

Experiments

- Datasets: Newswire collections where each collection is related to a major international event.
- Metrics: Average Precision@K and MAP (with paired t-test).
- Methods: **TF-IDF**, **TextRank**, **NERank**$_{Uni}$ (which assigns prior topic ranks uniformly), **NERank**$_{\alpha=0}$ (which sets $\alpha = 0$ in random walk process) and **NERank**$_{Full}$ (proposed approach).
- Results: **NERank**$_{Full}$ outperforms all the baselines.

Table 1: Experimental Results ($\star$: p-value$\leq$0.05)

Method	AvgP@5	AvgP@10	AvgP@15	MAP
TF-IDF	$0.85^\star$	$0.79^\star$	$0.73^\star$	$0.81^\star$
TextRank	$0.87^\star$	0.83	$0.73^\star$	$0.83^\star$
NERank$_{Uni}$	$0.80^\star$	$0.75^\star$	$0.71^\star$	$0.78^\star$
NERank$_{\alpha=0}$	$0.72^\star$	$0.61^\star$	$0.51^\star$	$0.62^\star$
NERank$_{Full}$	**0.92**	**0.87**	**0.79**	**0.89**

Conclusion and Future Work

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

References

[1] V. Jijkoun, M. A. Khalid, M. Marx, and M. de Rijke.
Named entity normalization in user generated content.
In *AND*, pages 23–30, 2008.

[2] R. Mihalcea and P. Tarau.
Textrank: Bringing order into text.
In *EMNLP*, pages 404–411, 2004.

[3] W. Shen, J. Wang, P. Luo, and M. Wang.
LINDEN: linking named entities with knowledge base via semantic knowledge.
In *WWW*, pages 449–458, 2012.

[4] G. B. Tran, M. Alrifai, and E. Herder.
Timeline summarization from relevant headlines.
In *ECIR*, pages 245–256, 2015.

[5] G. B. Tran, M. Alrifai, and D. Q. Nguyen.
Predicting relevant news events for timeline summaries.
In *WWW*, pages 91–92, 2013.