A Hierarchical Knowledge Representation for Expert Finding on Social Media

Yanran Li[1], Wenjie Li[1], and Sujian Li[2]

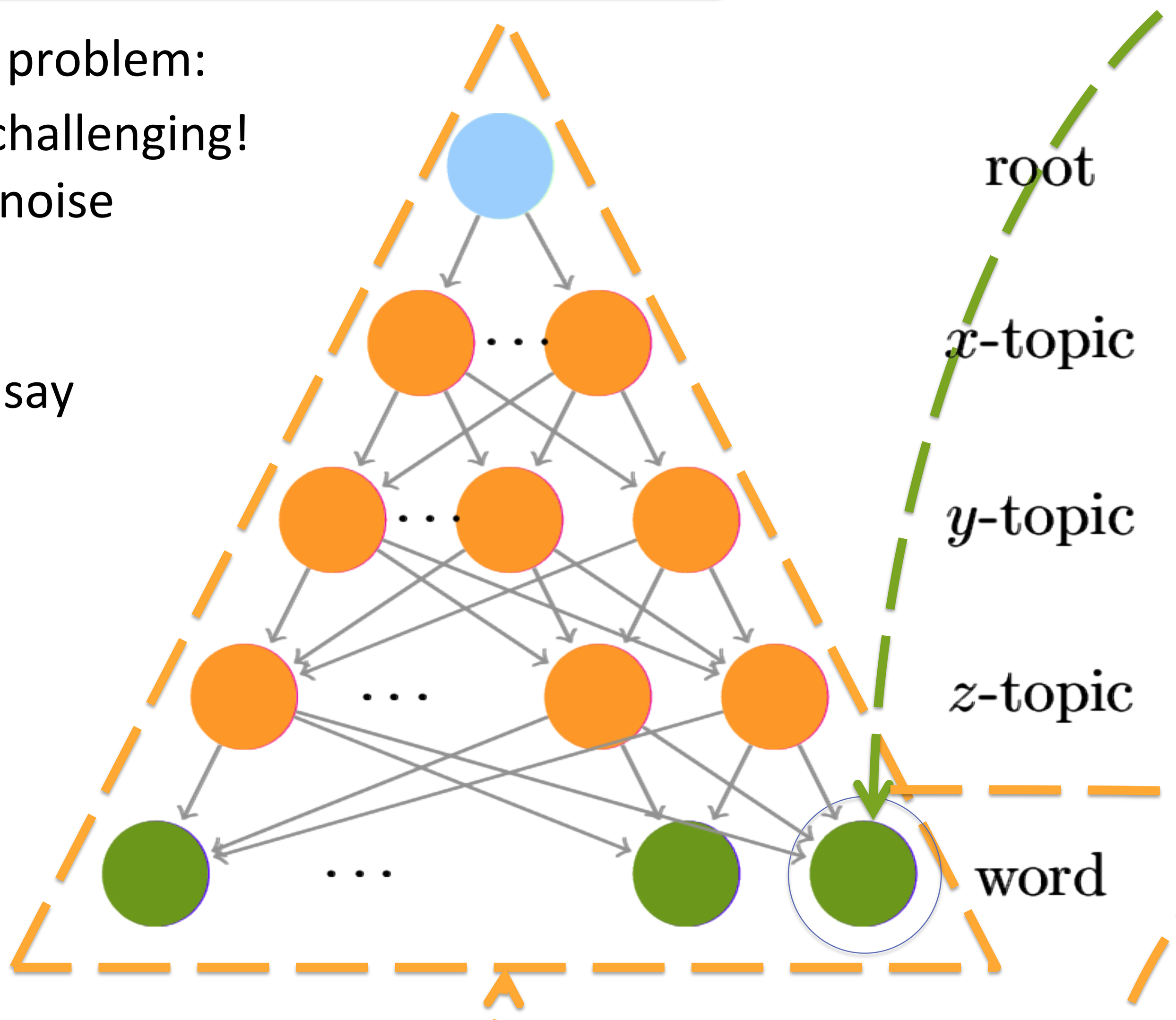
[1]Computing Department, Hong Kong Polytechnic University

[2]Key Laboratory of Computational Linguistics, Peking University

Semantic Matching for Expert Finding

We cast expert finding into matching problem:

- Expert Finding on Social Media is challenging!
  - Information on Social Media is noise
  - Expert ≠ Celebrity
  - Expert is **domain** specific
- Expert Knowledge is in What they say
  - Tweets
  - Retweets
- Knowledge is **Semantic**
  - Latent topic
- Knowledge is **HIERARCHICAL**
  - Generic to specific



Embedding for Tree Node

- Motivation
  - Words in the nodes are sparse
  - Contexts on Social Media are sparse
- Model
  - Skip-Gram in word2vec tool
- Calculation
  - Cosine similarity
  - Directly serve for approximate matching

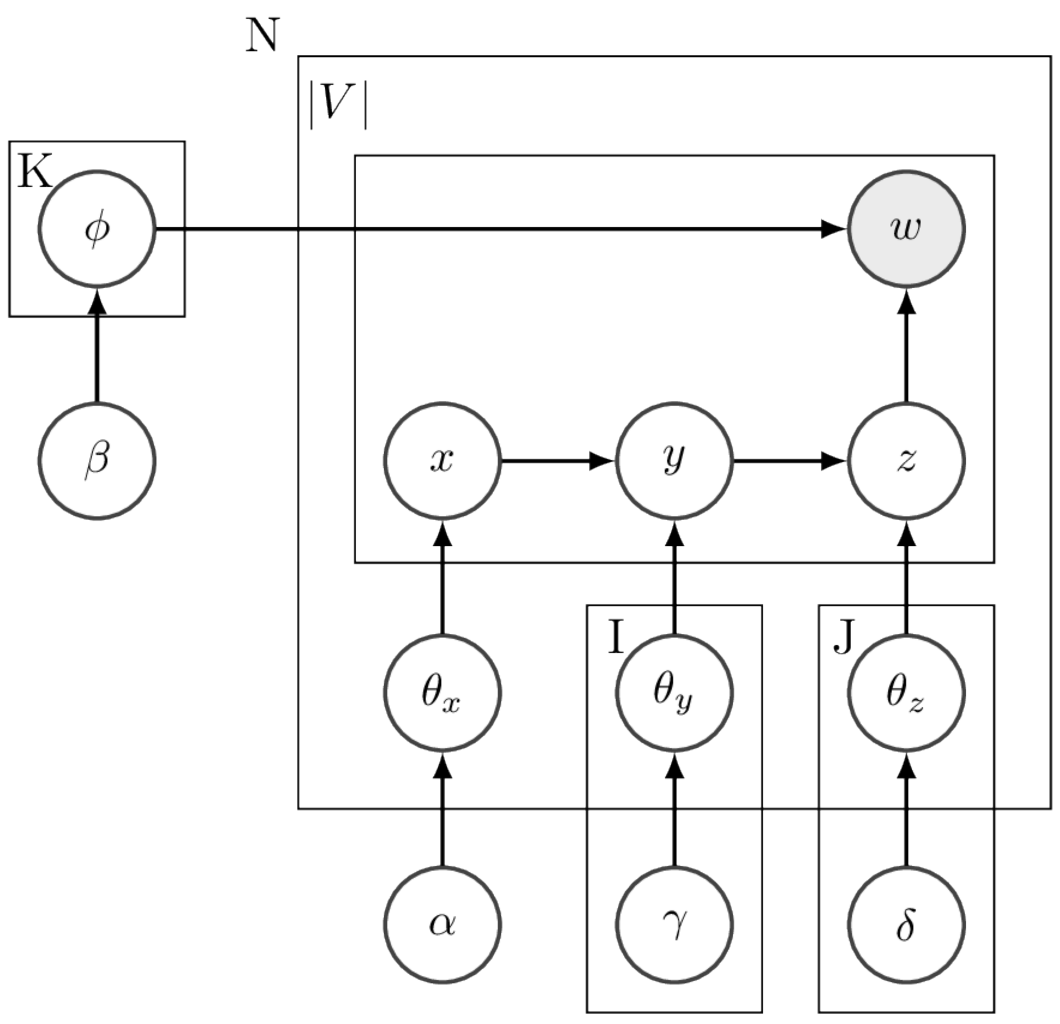
Approximate Tree Matching

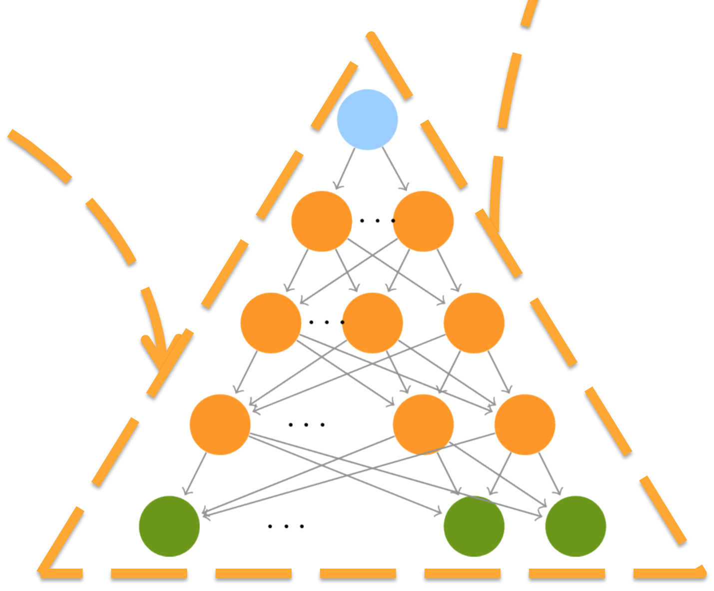
- Edit distance Based Matching
- Sum of the **Cost** of Editing Operation Sequence
- **3 Editing Operations:**
  - Substitution

$$\sigma(a \to b) = \begin{cases} 0, & a = b \\ \mathrm{sim}(a,b), & \mathrm{sim}(a,b) > 0.55 \\ \texttt{MAX\_VALUE}, & otherwise \end{cases}$$

  - Insertion $\texttt{MAX\_VALUE}$
  - Deletion $\texttt{MAX\_VALUE}$

Hierarchy for Knowledge Tree





- Pachinko Allocation Model
- Hierarchical Knowledge Tree
- For Each User
- For Each Domain

Topic Correlations:
LDA and other topic require that each topic should be independent with each other.
Too strict!
Instead, PAM can capture topic correlations.

Dataset and Experiments

- The experiments are conducted on 5 domains (i.e., *Beauty Blogger*, *Beauty Doctor*, *Parenting*, *E- Commerce*, and *Data Science* in Sina Microblog.
- For PAM:
  - Training: #113,924 posts from 40 experts in each domain.
  - Testing: 40 users randomly selected from the official expert lists as positive, 40 wrongly categorized users as negative.
  - Parameters: 5-level PAM, I=10, J=20, K=20.
- For Word Embedding:
  - Model: Skip-Gram
  - Training: another 25 million Sina Microblog posts and nearly 100 million tokens.
  - Parameters: 50 dimensions.

Approach	Precision		Recall		F-Score	
	Macro	Micro	Macro	Micro	Macro	Micro
unigram	0.380	0.484	0.615	0.380	0.469	0.432
bigram	0.435	0.537	0.615	0.435	0.507	0.486
LDA	0.430	0.473	0.540	0.430	0.474	0.451
Twitter-LDA	0.675	0.763	0.680	0.430	0.675	0.451
PAM	**0.720**	**0.818**	**0.720**	**0.720**	**0.714**	**0.769**

- In general, LDA, Twitter-LDA and PAM outperform unigram and bigram, showing the strength of latent semantic modeling.
- Our 5-level PAM gains observed improvement over Twitter-LDA.
  - Tree representation over vector space feature representation
  - Word embedding and partial matching
- The higher micro-recalls of PAM demonstrate its better generalization ability.

Conclusions

- **Hierarchy** is important!
- **Correlations** between topics is important!
- **Word embedding** well tackled sparseness!

- We formulate the expert finding task as a tree matching problem with the hierarchical knowledge representation.
- The experimental results demonstrate the advantage of using 5-level PAM and semantic enhancement against n-gram models and LDA-like models.
- It is flexible to incorporate more information to enrich the hierarchical representation.

Contact Us

{csyli, cswjli}@comp.poly.edu.hk
lisujian@pku.edu.cn

Scan Me To Download!