# Microsoft Research Asia at NTCIR-12 STC Task

Zhongxia Chen [2], Ruihua Song [1] and Xing Xie [1]

[1] Social Computing Group, Microsoft Research Asia

[2] University of Science and Technology of China

# Motivations

- Similar comment?
  - May be relevant but repetitive

| | |
|---|---|
| **Test Post** | 看这几天天气挺好的，到上海去玩一圈<br>Since the weather is good these days, I'd like to go to Shanghai for a visit |
| Comment1 | 到上海去玩?<br>Going to Shanghai? |

# Motivations

- Similar post?
  - Not bad

| | |
|---|---|
| **Test Post** | 看这几天天气挺好的，到上海去玩一圈<br>Since the weather is good these days, I'd like to go to Shanghai for a visit |
| Comment1 | 到上海去玩?<br>Going to Shanghai? |
| Similar Post | 看这几天天气挺好的，到北京去玩一圈<br>Since the weather is good these days, I'd like to go to Beijing for a visit |
| Comment2 | 羡慕<br>Envying you |

# Motivations

- Similar post?
  - But perhaps too specific

| | |
|---|---|
| **Test Post** | 看这几天天气挺好的，到上海去玩一圈<br>Since the weather is good these days, I'd like to go to Shanghai for a visit |
| Comment1 | 到上海去玩?<br>Going to Shanghai? |
| Similar Post | 看这几天天气挺好的，到北京去玩一圈<br>Since the weather is good these days, I'd like to go to Beijing for a visit |
| Comment2 | 羡慕<br>Envying you |
| Comment3 | 推荐去颐和园<br>I recommend you to go to the Summer Palace |

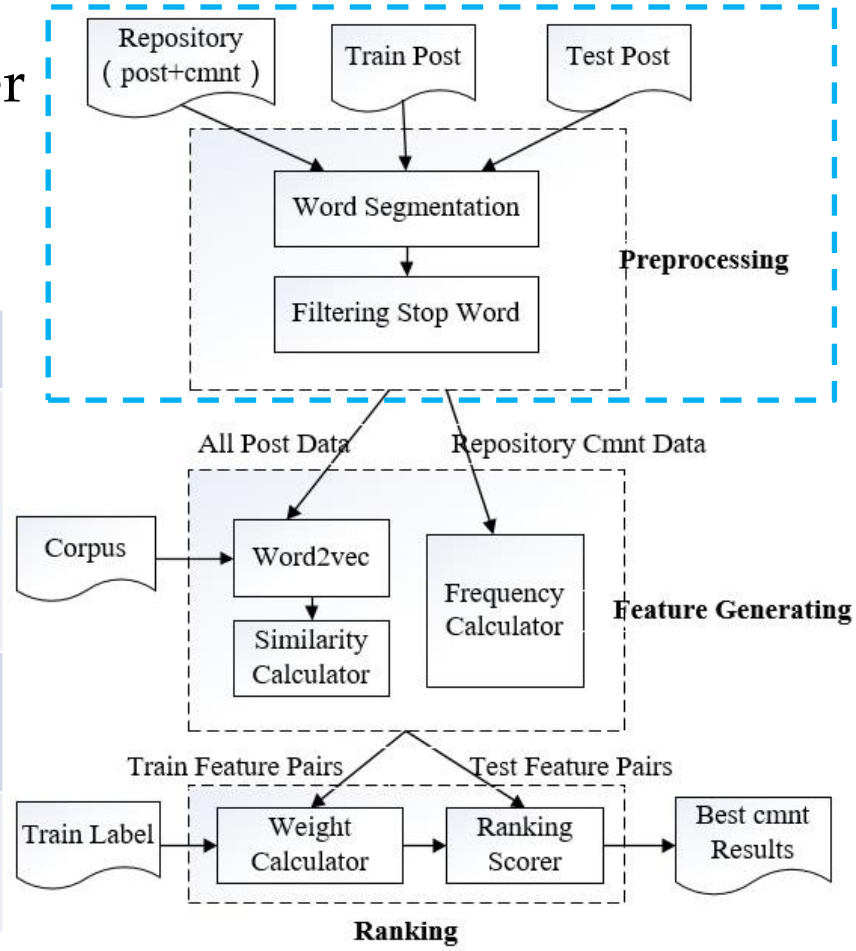
# Our Ideas

- Two observations
  - The comments responded to similar posts are good candidates
  - The more common a comment is, the less risky

# System Description

- Preprocessing
  - Stanford Chinese Word Segmenter
  - Filter stop words

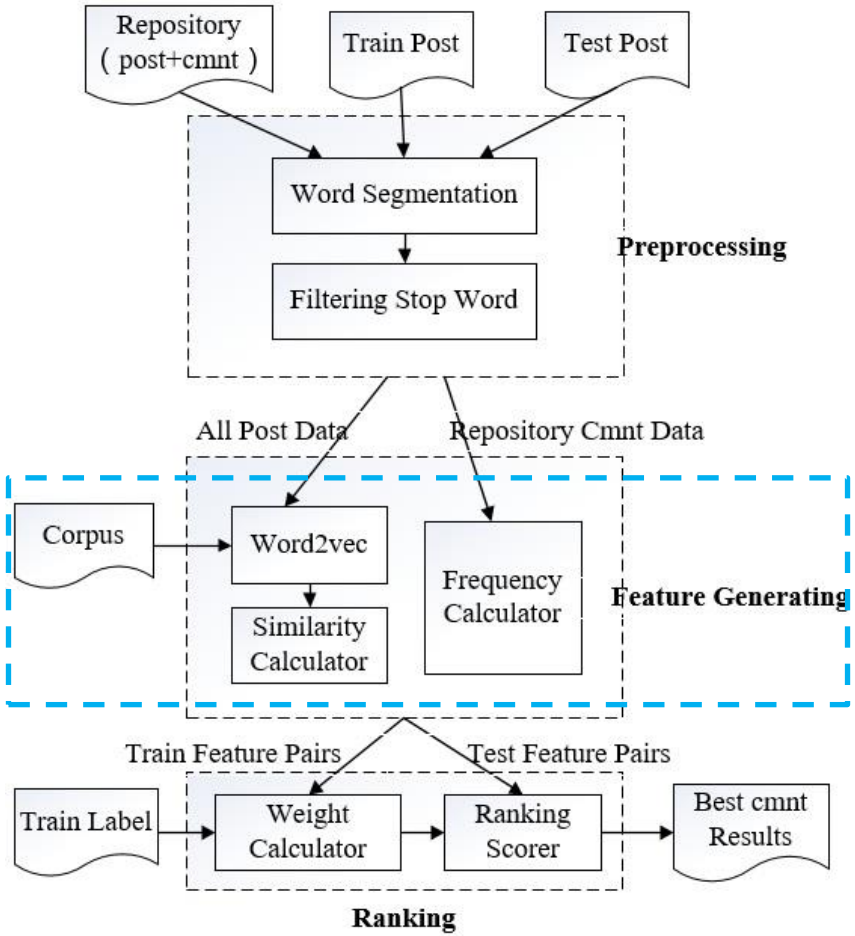
| | |
|---|---|
| Short Text ID | repos-cmnt-1000578260 |
| Origin Text | 西班牙踢球多黏啊总看传球了’意大利必胜<br>(Spain always passes and I will support Italy) |
| Segmentation Result | 西班牙 踢球 多 黏 啊 总 看 传球 了 ’ 意大利 必胜 |
| Filtering Result | 西班牙 踢球 黏 总 传球 意大利 必胜 |



# System Description

- Feature Generation
  - Post-Post Similarity
    - Generate word vectors
      - Google Word2Vec
      - External Chinese Corpus
    - Generate short text vectors
      - Summation of word vectors
    - Calculate similarity
      - Cosine similarity

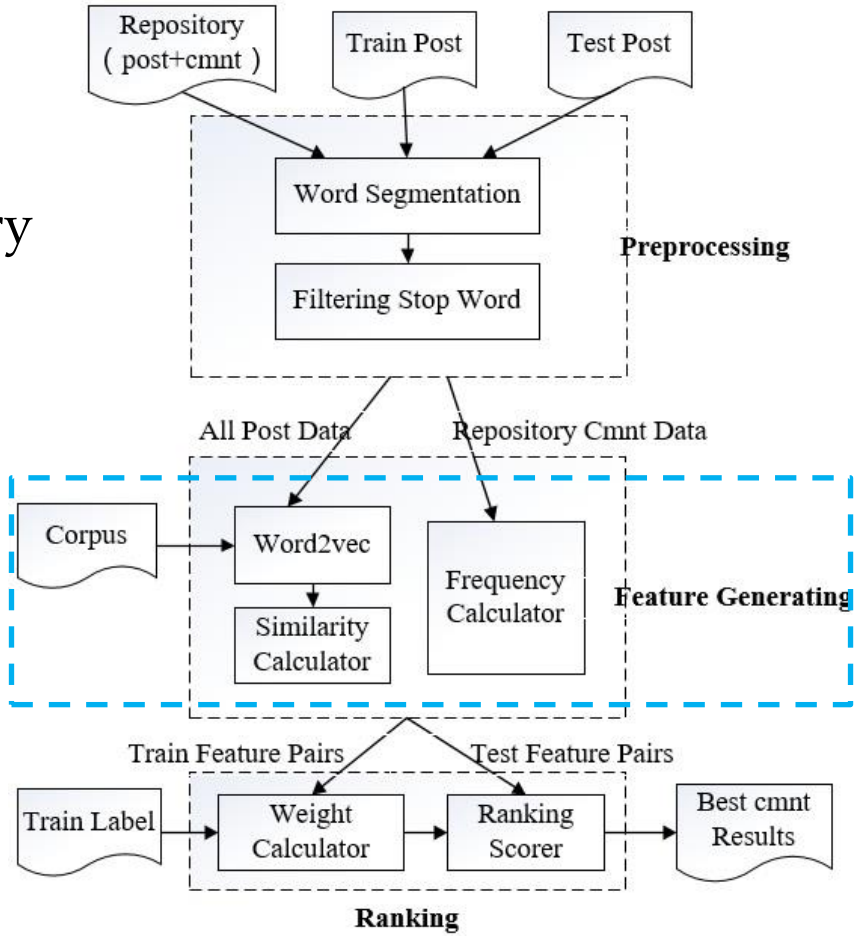
$$F_{sim}(p, Ci) = \frac{\overrightarrow{v_p} \cdot \overrightarrow{v_{Ci}}}{|\overrightarrow{v_p}| \cdot |\overrightarrow{v_{Ci}}|}$$



# System Description

- Feature Generation
  - Popularity
    - Count frequencies of comments
      - Top ten comments in repository

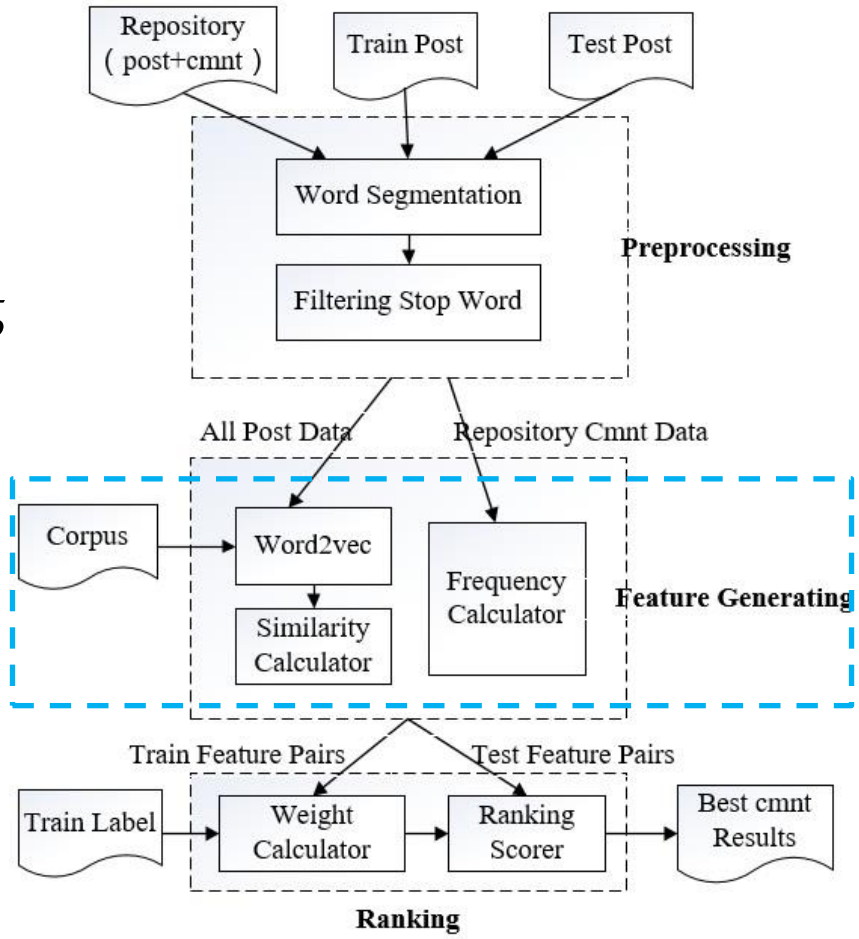
| Comment | Frequency |
|---|---|
| 哈哈(laugh) | 14830 |
| 哈哈哈(laugh) | 8096 |
| 呵呵(no comment) | 8075 |
| 嗯(fine) | 4443 |
| 哈哈哈哈(laugh) | 4297 |
| 不错(not bad) | 4222 |
| 好(good) | 3676 |
| 是的(yes) | 3472 |
| 喜欢(I like it) | 3216 |
| 赞(great) | 3209 |



# System Description

- Feature Generation
  - Popularity
    - Count frequencies of comments
    - Considering Power Law
    - Generate popularity feature by log function of frequency

$$F_{pop}(Ci) = \log(Frequency(Ci))$$

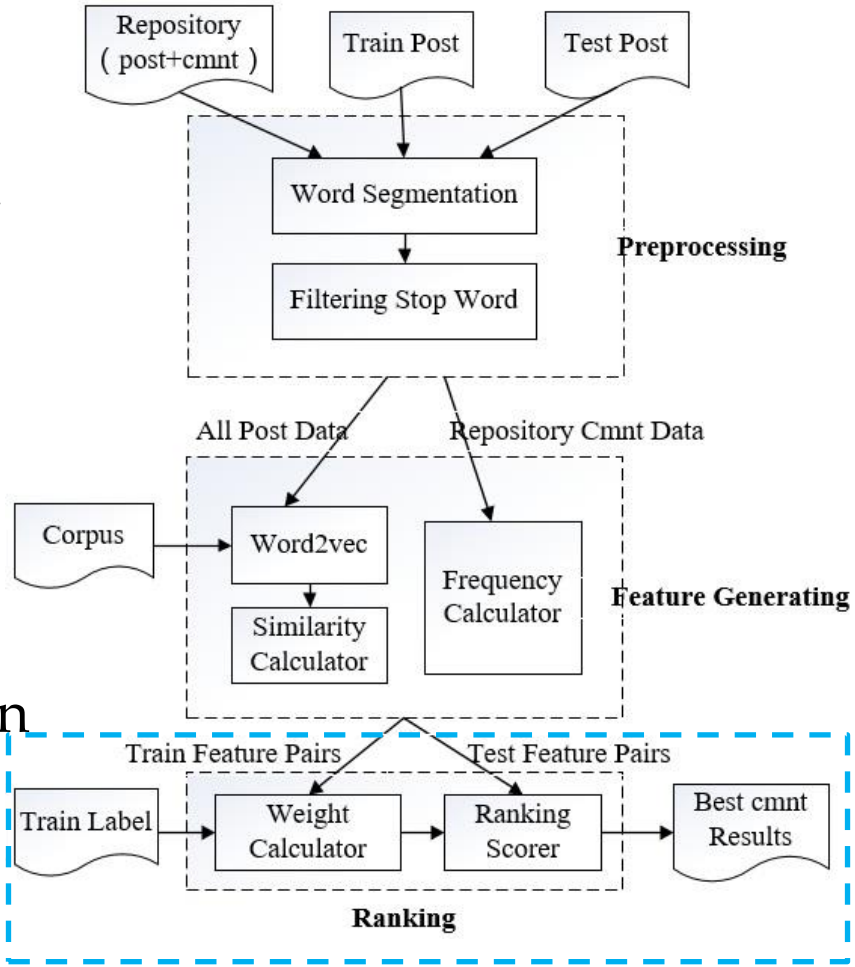


# System Description

- Ranking
  - Retrieve top k comments with closed corresponding posts using similarity feature, then re-rank by popularity in a small range
  - Train a linear model

$$Score(p, Ci) = F_{sim}(p, p') + w \cdot F_{pop}(Ci)$$

where, Ci is a comment of p'. Then rank by score



# Experiments

- Submitted three runs
  - MSRSC-C-R1: ranking by a linear combination of post-post similarity and popularity
  - MSRSC-C-R2: retrieving top 50 comments by post-post similarity and re-ranking by popularity
  - MSRSC-C-R3: ranking by post-post similarity
- Word2vec settings
  - Skip-gram model
  - Window size is 10
  - Vector length is 100

# Experiments

- Result for 3 runs

| Run | Mean nDCG@1 | Mean P+ | Mean nERR@10 |
|---|---|---|---|
| **MSRSC-C-R1** | **0.3367** | **0.4854** | **0.4592** |
| MSRSC-C-R2 | 0.2733 | 0.4208 | 0.3857 |
| MSRSC-C-R3 | 0.0933 | 0.2420 | 0.2236 |

  - Population is surprisingly helpful
  - Linear model is better than retrieval and re-ranking model
    - Range of best fifty similar posts is uncertain

# Conclusions

- Post-post similarity is useful
  - We use word embedding in representing short texts
- Popularity of comments is helpful when being combined with post-post similarity

# Thank you!

Contact: rsong@microsoft.com