Nders at the NTCIR-13 Short Text Conversation 2 Task

Han Ni, Liansheng Lin, Ge Xu
NetDragon Websoft Inc.

网龙网络公司
NETDRAGON WEBSOFT INC.

Introduction

This poster describes our retrieval-based approaches at NTCIR-13 short text conversation 2 (STC-2) task (Chinese).
For a new post, our system firstly retrieves similar posts in the repository to get their corresponding comments, and then finds the related comments directly from the repository.
Moreover, we devise two new methods:
1) LSTM-Sen2Vec model to get the vector of sentence.
2) Pattern-IDF to rank the candidates from above.
Our best run achieves 0.4780 for mean nG@1, 0.5497 for mean P+, and 0.5882 for mean nERR@10, and respectively rankes 4th, 5th, 5th among 22 teams.

Example from Our Results

Post	好喜欢这种手绘的画啊，小柚子加油 Really like this hand-painted painting, xiaoyouzi, come on!
Comment 1	喜欢这种手绘我也要画 Like this hand-painted painting, I also want to draw it.
Comment 2	我想问你的画是手绘还是电脑画的啊 I want to ask whether your paintings are hand-painted or computer painted.
Comment 3	这是用电脑画的还是手绘的? Was this drawn by computer or hand-painted ?
Comment 4	很有爱的画。是用手绘板画的么? It's a lovely painting. Was it drawn with a hand-painted board?

Our System

Preprocessing

- Convert traditional Chinese into simplified Chinese
- Convert full-width characters into half-width ones
- Word Segmentation (PKU standard)
- Replace all the number, datetime, url with token _NUM, _TIME, _URL respectively
- Filter meaningless words and symbols

Candidates Generation

- Train TF-IDF, Word2Vec, LSA, LDA and LSTM-Sen2Vec models on the repository and convert sentence into vector
- Use cosine similarity to calculate similarity between two vectors
- Retrieve similar posts with Word2Vec, LSA/LDA and LSTM-Sen2Vec models
- Get corresponding comments from the similar posts as part of comment candidates
- Retrieve appropriate comments directly from all comments with Word2Vec and LSA/LDA models
- Combine two candidates as the final comment candidates

Ranking

- Use graph-based algorithm TextRank to rank the candidates
- Use Pattern-IDF to rank the candidates
- Use Pattern-IDF and TextRank to rank the candidates

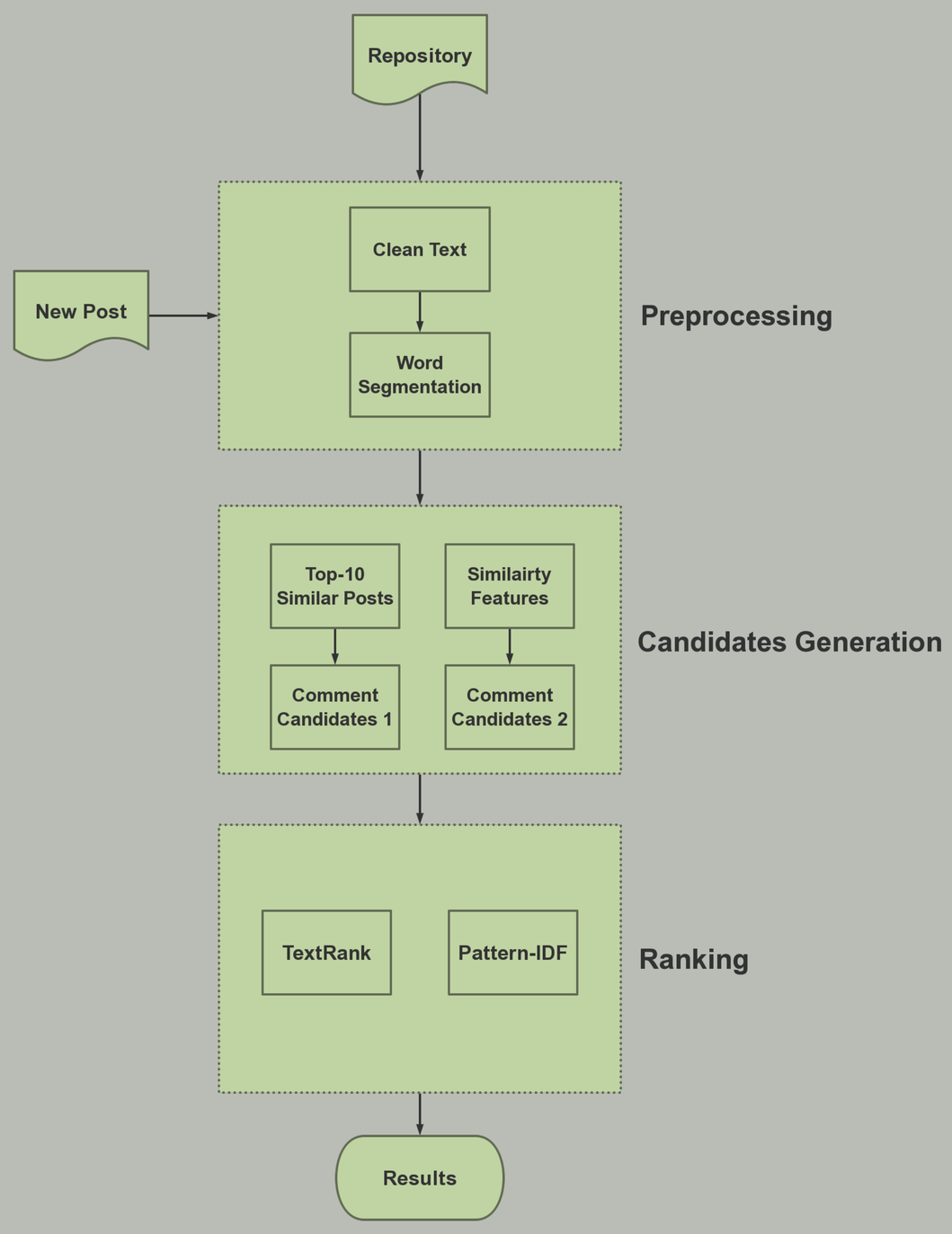

Figure 1: System Architecture

Experimental Results

Run	Mean nG@1	Mean P+	Mean nERR@10
Nders-C-R1	0.4593	0.5394	0.5805
Nders-C-R2	0.4743	**0.5497**	**0.5882**
Nders-C-R3	0.4647	0.5317	0.5768
Nders-C-R4	**0.4780**	0.5338	0.5809
Nders-C-R5	0.4550	0.5495	0.5868
R2 vs. R4	↓0.77%	↑1.26%	↑2.02%

- Nders-C-R5: LDA + Word2Vec + LSTM-Sen2Vec
- Nders-C-R4: LSA + Word2Vec + LSTM-Sen2Vec
- Nders-C-R3: R4 + TextRank (words as vertices)
- Nders-C-R2: R4 + Pattern-IDF
- Nders-C-R1: R4 + Pattern-IDF + TextRank (sentences as vertices)

Conclusions

We propose an approach for STC-2 task of NTCIR-13. The LSA, Word2Vec and LSTM-Sen2Vec model are used to find similar posts. The LSA and Word2Vec model are used to retrieve comment candidates. A graph-based algorithm TextRank and the Pattern-IDF we devised are applied to rank the candidates. Results show that the Pattern-IDF we devised is effective for ranking while TextRank not, and LDA model outperforms LSA model in retrieving candidates.

Acknowledgement

NetDragon Websoft Inc.
NTCIR-13 STC-2 Organizers

{nihan, linliansheng, XuGeNLP}@nd.com.cn