SJTU SPEECH LAB

上海交通大學智能語音實驗室

The splab at the NTCIR-12 STC Task

Ke Wu, Xuan Liu and Kai Yu

Emails: { ke.wu, 1097114522, kai.yu }@sjtu.edu.cn

Motivation

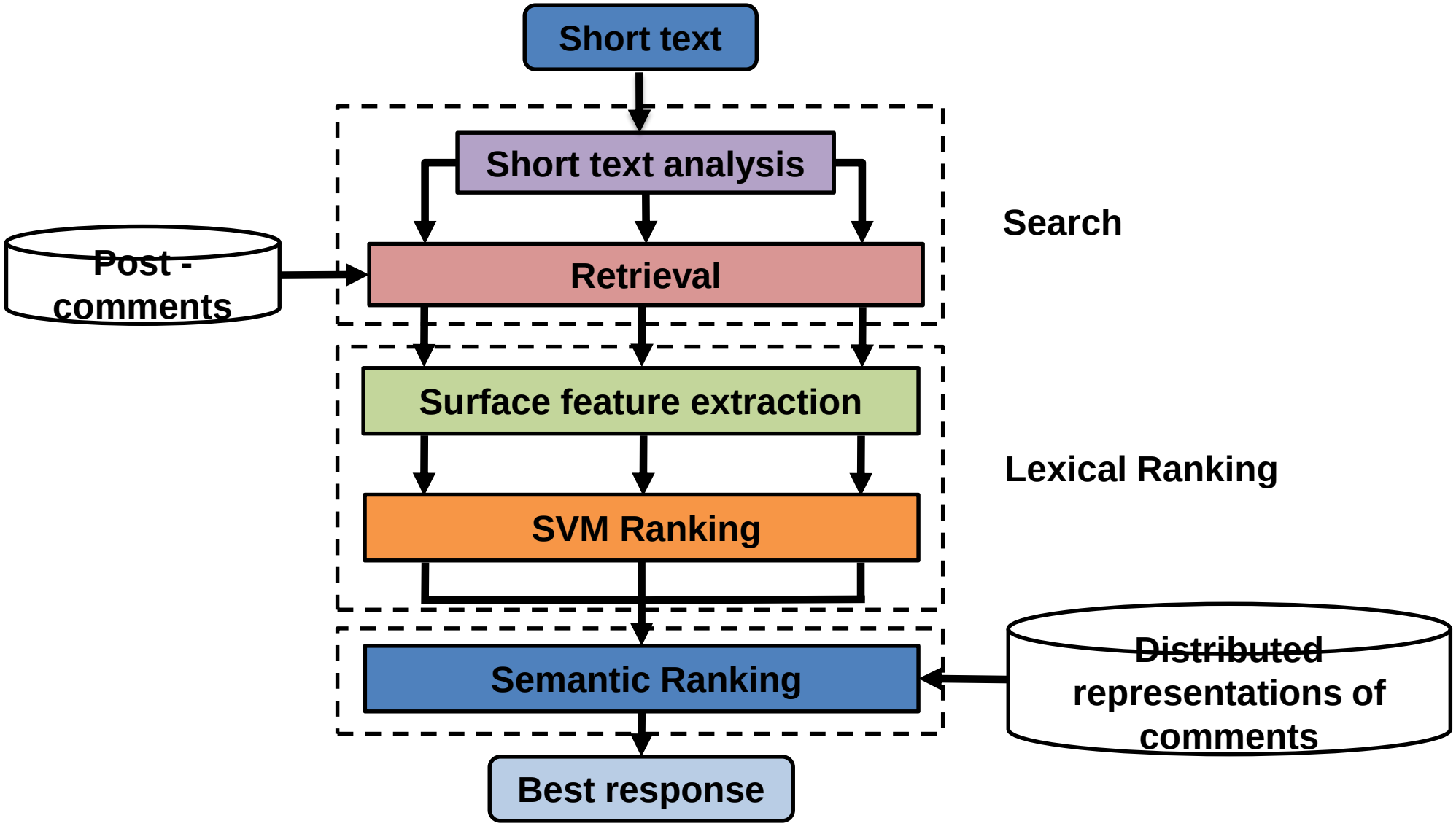
- Simlar to finding a needle in a haystack, it is hard to obtain a proper response in an exetremely large conversation data

- Too many irrelevant candidate comments can potentially hinder a system's ability to identify the appropriate response from the large pool of candidates

Basic Idea

- To collect suffcient candidate comments including the suitable responses in the small candidate pool.
- Our system attempts to facilitate high short text conversation performance by a three-level ranking framework.
  - Through a couple of selection turns, we generate the final plausible candidate set.
  - In the small pool of candidates, we leverage deep learning techniques to find the best response.

Our three-tier ranking framework



Search Component

- Short text analysis
  - Tailor a short text to use IR technologies
  - Translate a short text into multiple terms with weights

- Retrieval
  - To index all post-comment pairs for a highly-efficient information filtering

Short text analysis

- Method 1: MG
  - all potential words with equal weights
  - to improve search recall
- Method 2: TFIDF
  - keywords by TF-IDF
- Method 3: TextRank
  - Similar to Google's PageRank algorithm

Detect or recognize the focus with salient information in a short text

Retrieval

- To index all post-comments pairs (offline)

* To use the default similarity function in Lucene (online)

$$\text{score}(q,s) = \text{co}(q,s)\cdot\text{qn}(q)\cdot\sum_{t\in q}\{\text{tf}(t\in s)\cdot\text{idf}(t)^2\cdot w_t\cdot\text{norm}(t,s)\}$$

Lexical Ranking Component

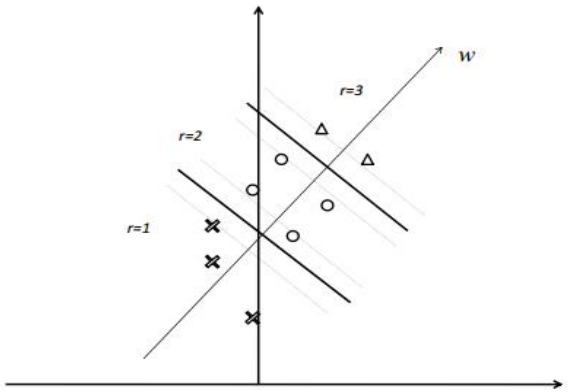
Goal: To try to promote all the relevant comments to the top of a ranked list based on the downstream comment candidates by leveraging a small portion of lablled post-comment pairs.

Lexical Features:

- Similarity features
- Matching features
  - Longest common string and co-occurring statistics

Ranking SVM:

Used to exploit lexical features to ranking the candidates



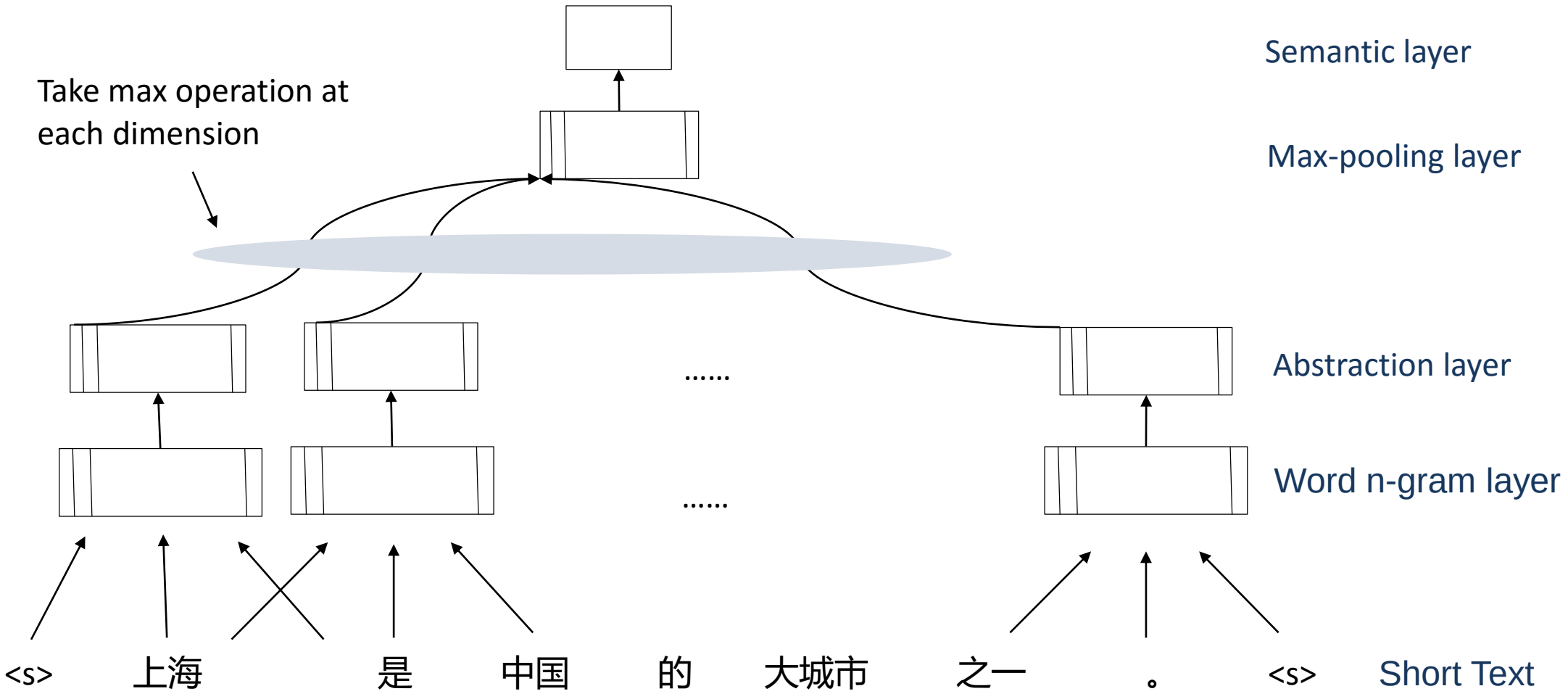
Lexical Features

Types	Features	Meanings
Similarity Features	Q2C	Similarity between the query q and the candidate comment c
	Q2P_Ave	Average of the similarities between the query q and the posts with which the candidate comment c is paired
	Q2P_Max	Maximum of the similarities between the query q and the posts with which the candidate comment c is paired
	Q2P_Min	Minimum of the similarities between the query q and the posts with which the candidate comment c is paired
Matching Features	LCS	Length of the longest common string between the query q and the candidate comment c
	LCS_Rate	Ratio of LCS to the length of the candidate comment
	Co_Size	Number of co-occuring words between the query q and the candidate comment c
	Co_Rate	Ratio of Co_Size to the number of words in the candidate comment c
	Co_IDF_Sum	Sum of IDFs of co-occuring words between the query q and the candidate comment c
	Co_IDF_Ave	Average of IDFs of co-occuring words between the query q and the candidate comment c
Others	Post_Num	Number of the posts with which the candidate comment c is paired

Semantic Ranking Component

- Receive and re-rank the aggregated results of three different search strategies from the lexical ranking component by semantics

Our Sentence Embedding



Submitted Results

Run name	Mean nDCG@1	Mean P+	Mean nERR@10
splab-C-R1	0.2933	0.4735	0.4449
splab-C-R2	0.0967	0.2069	0.1831
splab-C-R3	0.0967	0.1896	0.1650

- splab-C-R1 uses the three-tie framework
- splab-C-R2 uses the first method on RHS
- splab-C-R3 uses the third method on RHS

No.	weight update	post vector	cmnt vector
1	asynchronous	L	R
2	asynchronous	R	R
3	shared	L or R	L or R

Conclusions

- we described our system's three-pronged strategy for identifying proper responses that balance high candidate recall and processing time for candidate scoring.
- The evaluation on a test set of 100 test queries provided by the organizers shows that our three-tier ranking system is effective.

THANK YOU
for your
ATTENTION!