NTCIR-13, December 2017, Tokyo, Japan

SG01 at the NTCIR-13 STC-2 task

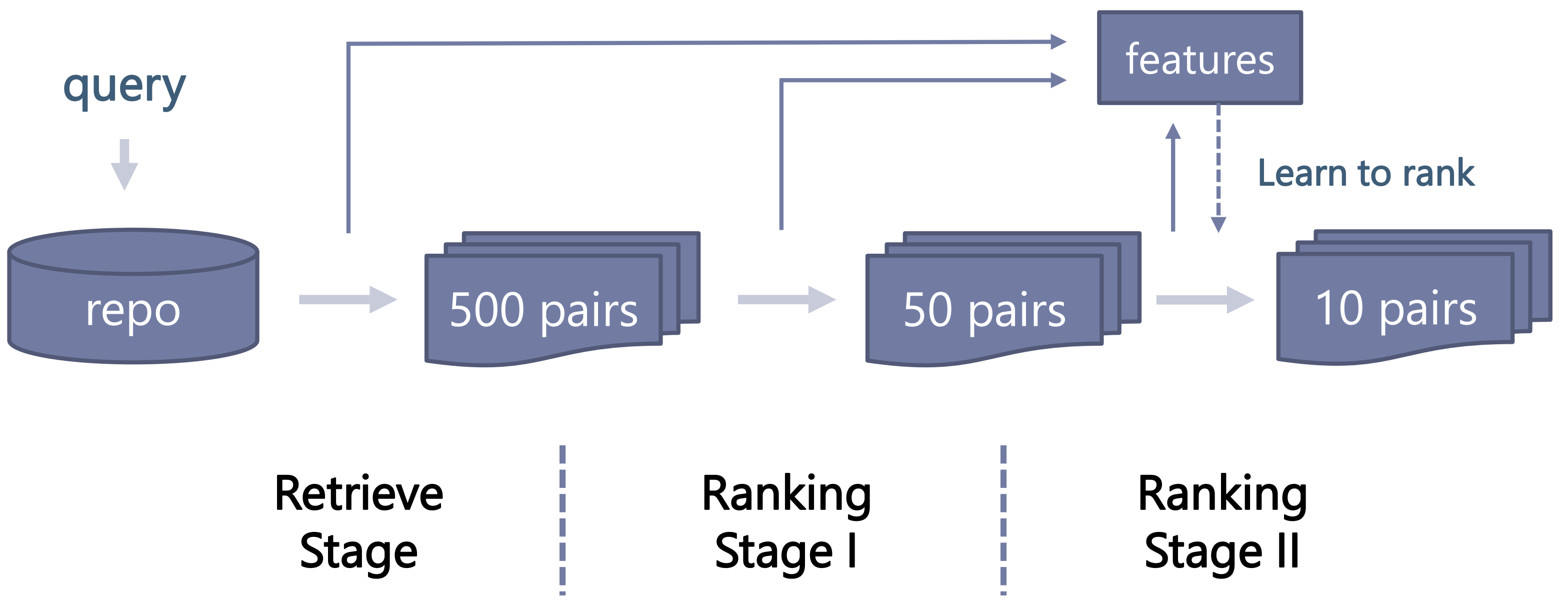
Haizhou Zhao, *Yi Du, Hangyu Li,*
Qiao Qian, Hao Zhou,
Minlie Huang, Jingfang Xu

Sogou Inc. | Beijing, China
Tsinghua University | Beijing, China

Introduction

- Team Name: ***SG01***
  - Joint Team from **Sogou Inc.** and **Tsinghua University**
- Subtask: ***Chinese subtask***
  - Retrieval-based method: 3 submissions
  - Generation-based method: 5 submissions
  - **Top performance** in both methods
- Next …
  - Retrieval-based Method
  - Generation-based Method
  - Conclusions
  - Q & A

Overview of Retrieval-based Method



Retrieve Stage | Retrieval-based Method

- Data-Preprocessing
  - Remove ***frequent, advertising*** and ***short*** post-comment pairs
- Put the repository into a light-weighted search engine
  - Treat post-comment pairs as webpages
- Retrieve **500 pairs** for a given query (or “new post”)
- Keep the calculated features for searching for later usage
  - BM25
  - MRF for term dependency [D. Metzler 2005]
  - Proximity [T. Tao 2007]
  - …

Ranking Stage I | Retrieval-based Method

- Employ features more intuitive in STC task
  - **cosine similarity** of TF-IDF vector between …
  - negative **Word Mover Distance** [M. J. Kusner 2015] between …
    - query ↔ post
    - query ↔ comment
    - query ↔ post + comment
  - **Translation based language model** [Z. Ji 2014] $Score_{trans}$
- Ranking
  - Treat each feature as a ranker
  - Simply add the sequence numbers to get a final rank
  - Keep **top 50 pairs**

Ranking Stage II: new features | Retrieval-based Method

- Employ more neural network features capturing richer structure in STC
  - $Score_{embd}$
  - $Score_{BiLSTM+CNN}$ [R. Yan 2016]

$$L = \max(0, 1 - s(x, y^{+}) + s(x, y^{-}))$$

↑ *Trained with **ranking based objective**, using given repository plus extra 12 million crawled post-comment pairs, noted as $Repo_{extn}$*

  - $Score_{S2S-p2c}$
  - $Score_{S2S-c2p}$

← *Defined later in Generation-based Method*

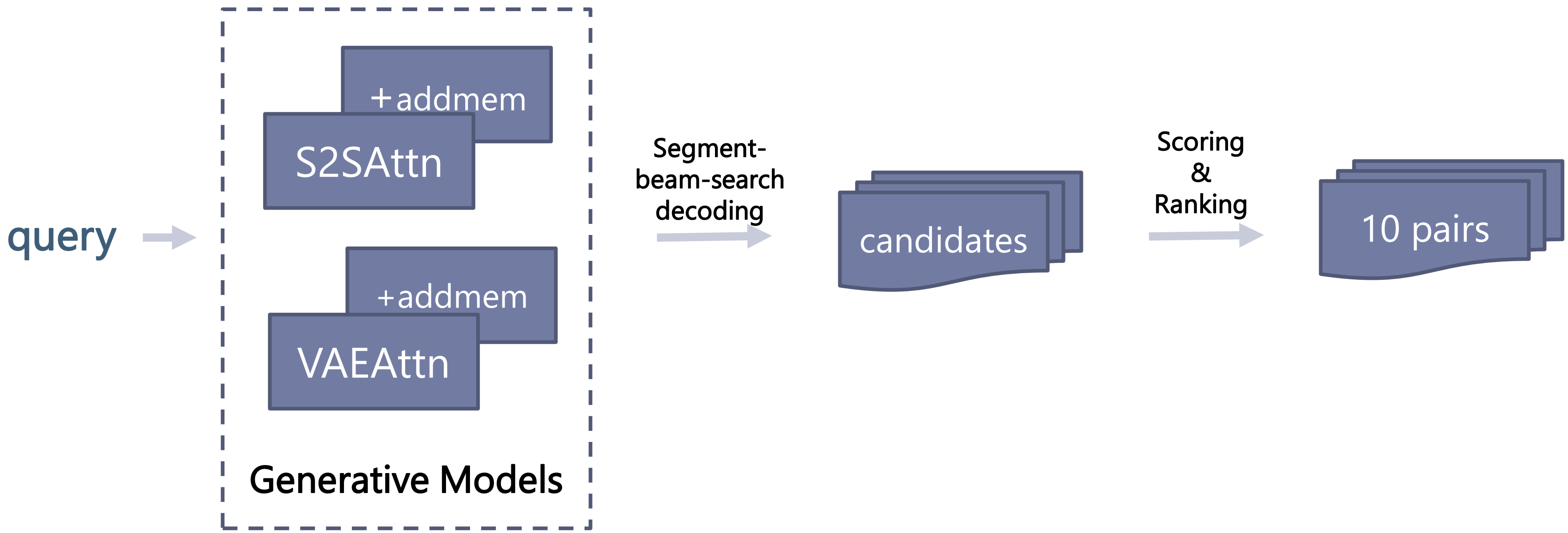
Ranking Stage II: learning 2 rank | Retrieval-based Method

- Use **all features aforementioned**
- Training data: given 11 thous. plus 30 thous. labeled pairs
- ***LambdaMART***
- **Top 10** to be the final result
- $Score_{trans}$ and $Score_{BiLSTM+CNN}$ are a little more important

Experiments | Retrieval-based Method

Submission	Learning to rank respect to which measure on training data	nG@1	P+	nERR@10
SG01-C-R1	nG@1	**0.5355**	0.6084	0.6579
SG01-C-R2	nERR@10	0.5168	0.5944	0.6461
SG01-C-R3	P+	0.5048	**0.6200**	**0.6663**

Overview of Generation-based Method



Generative Models | Generation-based Method

- $S2SAttn$
  - seq2seq [I. Sutskever 2014] with attention mechanism
- $S2SAttn{-}addmem$
  - Add dynamic memory to the attention
- $VAEAttn$
  - Use Variational Auto-Encoder
- $VAEAttn{-}addmem$

- Training data: $Repo_{extn}$ with data-preprocessing
- Decode using ***segment-beam-search***

Candidates Ranking: scores | Generation-based Method

- Scoring Features
  - ***likelihood***
    - $\log(P(Y'|X))$, for post X and generated comment Y'
    - We note score from one model as $Score_{s2s-p2c}$
    - For scores from different models (except VAE models) and implementations, we add them up as $\boldsymbol{Li}$
  - ***posterior***
    - $\log(P(X|Y'))$
    - $Score_{s2s-c2p}$
    - $\boldsymbol{Po}$
  - Calculated by our **well trained models**

Candidates Ranking: rank & output

- Ranking
  - $score = \frac{\lambda * Li + (1-\lambda) * Po}{lp(Y')}$
  - Discount factor $lp(Y') = \frac{(c+|Y'|)^{\alpha}}{(c+1)^{\alpha}}$ [Y. Wu 2016]
- Before Final Output: Process candidates by rules
  - Abandon candidates with keywords in blacklist
  - De-duplicate **consecutively repeated segments**
  - Truncate **consecutively repeated punctuations**

Experiments | Generation-based Method

Submission	Fusion of candidates from*	Scoring By**	nG@1	P+	nERR@10
SG01-C-G5	$VAEAttn, VAEAttn{-}addmem$	Li	0.3820	0.5068	0.5596
SG01-C-G4	$S2SAttn, S2SAttn{-}addmem$	Li	0.4483	0.5545	0.6129
SG01-C-G3	$S2SAttn, S2SAttn{-}addmem$	Li & Po	0.5633	0.6567	0.6947
SG01-C-G2	$VAEAttn, VAEAttn{-}addmem$	Li & Po	0.5483	0.6335	0.6783
SG01-C-G1	*All 4 kinds of models*	Li & Po	**0.5867**	**0.6670**	**0.7095**

**: could be multiple implementations for one model, using different subset of corpus and hyper-parameters*

***: all scores are discounted by lp*

Analysis | Generation-based Method

- The feature $\boldsymbol{Po}$ brings advantage **with statistical significance** to those without Po , by giving higher rank to more informative candidates
- $\boldsymbol{VAE}$ does worse than traditional seq2seq, but it can bring in interesting candidates
- Using **fusion of results from models** do better than relying on single model, because the ranking will bring preferable candidates to top 10

Conclusions

- Comparison between methods
  - **Generation-based method** does better, however, it still tends to generate "safe" responses
  - Retrieval-based method tends to get context-dependent or in-coherent comments
- Size of training data maters

References

- **Z. Ji**, Z. Lu, and H. Li. An information retrieval approach to short text conversation. CoRR, abs/1408.6988, **2014**.
- **M. J. Kusner**, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 957–966. JMLR.org, **2015**.
- **D. Metzler** and W. B. Croft. A markov random field model for term dependencies. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pages 472–479, New York, NY, USA, **2005**. ACM.
- **I. Sutskever**, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc., **2014**.
- **T. Tao** and C. Zhai. An exploration of proximity measures in information retrieval. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pages 295–302, New York, NY, USA, **2007**. ACM.
- **Y. Wu**, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144, **2016**.
- **R. Yan**, Y. Song, X. Zhou, and H. Wu. "Shall I Be Your Chat Companion?": Towards an Online Human-Computer Conversation System. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, pages 649–658, New York, NY, USA, **2016**. ACM.

Thank you!

zhaohaizhou@sogou-inc.com