Unsupervised Extractive Summarization Using Sparse Coding

UNC NLP

Somnath Basu Roy Chowdhury, Chao Zhao and Snigdha Chaturvedi
{somnath, zhaochao, snigdha}@cs.unc.edu
UNC Chapel Hill

brcsomnath/SemAE

Introduction

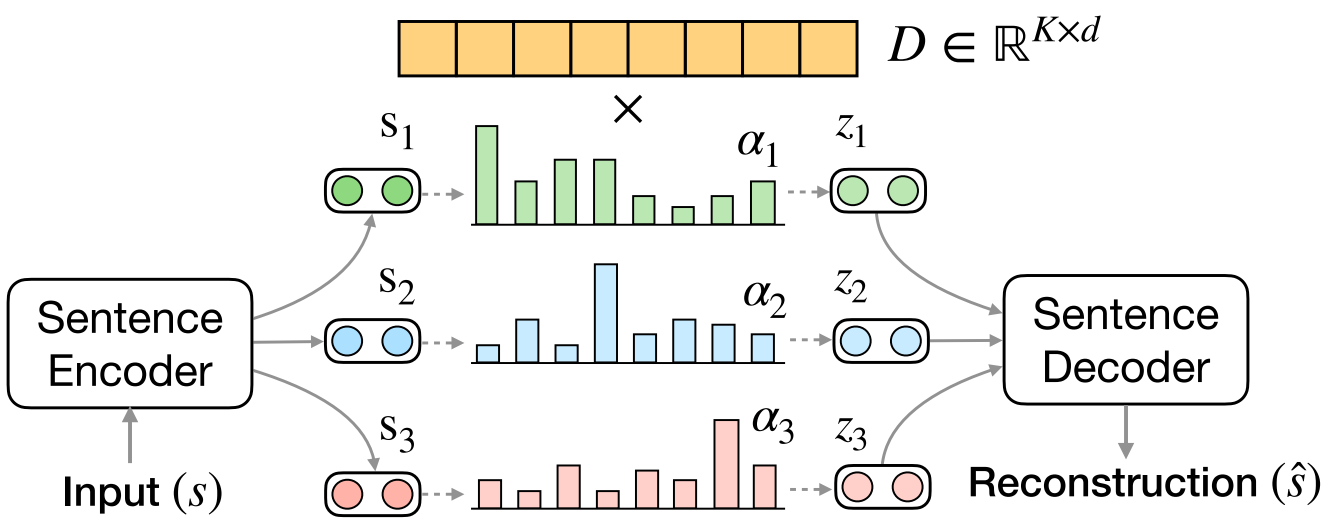
- Automatic opinion summaries enable faster comparison, search, and better consumer feedback understanding
- Unsupervised opinion systems are desirable due to the scarcity of labeled data
- It is important to understand the underlying semantics in an opinion
- The underlying semantics can be captured as a distribution over latent semantic units
- Opinions aligning with popular semantic distribution are selected to form the summary

Semantic Autoencoder (SemAE)

SemAE performs extractive opinion summarization in the following phases:

- Text Representation Learning
- Summarization based on saliency scores
  - General Summarization — relevance, redundancy and aspect-awareness
  - Aspect Summarization — relevance and informativeness

Representation Learning



- Encoder takes Input (s) to generate a multi-head sentence representation $[s_h]_{h=1}^{H}$
- A latent representation is constructed over the learnable dictionary $\alpha_h = \text{softmax}(s_h D^T)$
- The reconstructed vector $z_h = \alpha_h D$ is forwarded to the decoder to generate $\hat{s}$
- The model is trained to optimize the following loss:

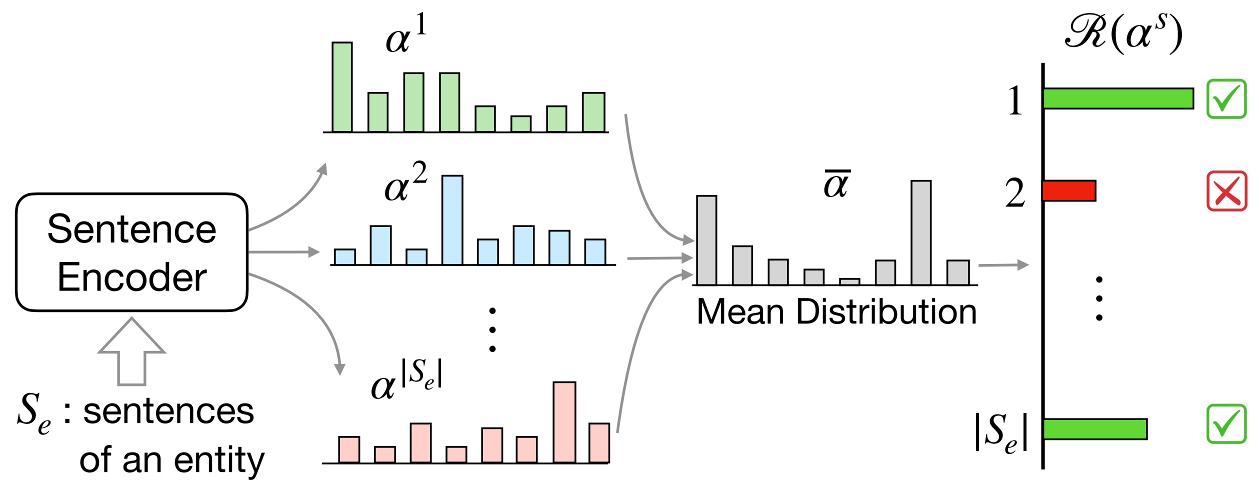
$$\mathscr{L}_{\text{CE}}(s, \hat{s}) + \lambda_1 \sum_h |\alpha_h| + \lambda_2 \sum_h H(\alpha_h)$$

Sentence Selection

Sentences are selected based on their saliency scores $\mathscr{R}(\alpha^s)$. $\mathscr{R}(\alpha^s)$ is computed using:

- Relevance: $\Delta(\bar{\alpha}, \alpha^s)$
- Redundancy: $-\gamma \max_{s' \in \hat{O}_e} \Delta(\alpha^{s'}, \alpha^s)$
- Aspect-awareness: Iterate over aspects and select salient sentences
- Informativeness: $-\beta\Delta(\alpha^B, \alpha^s)$

Summarization



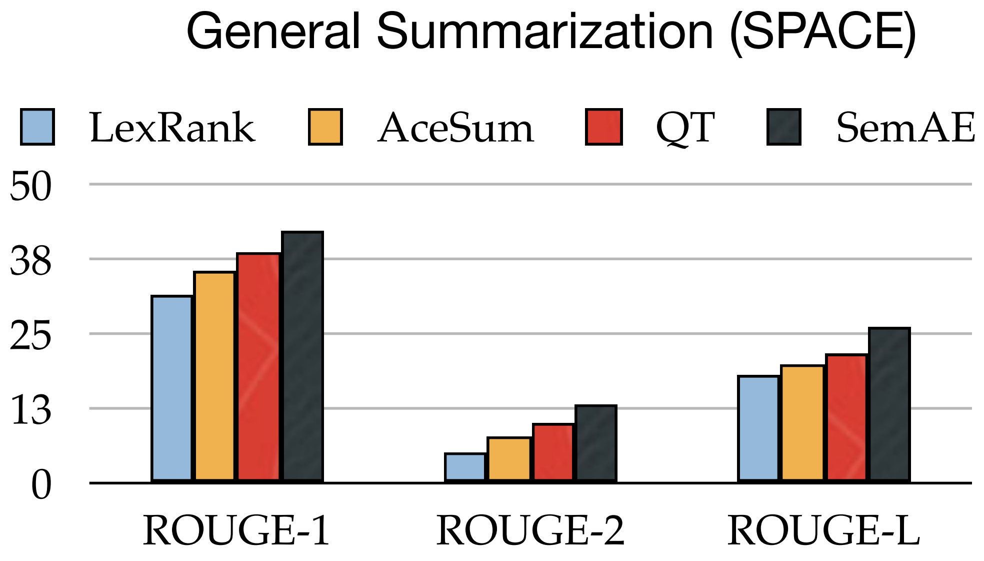
General Summarization:

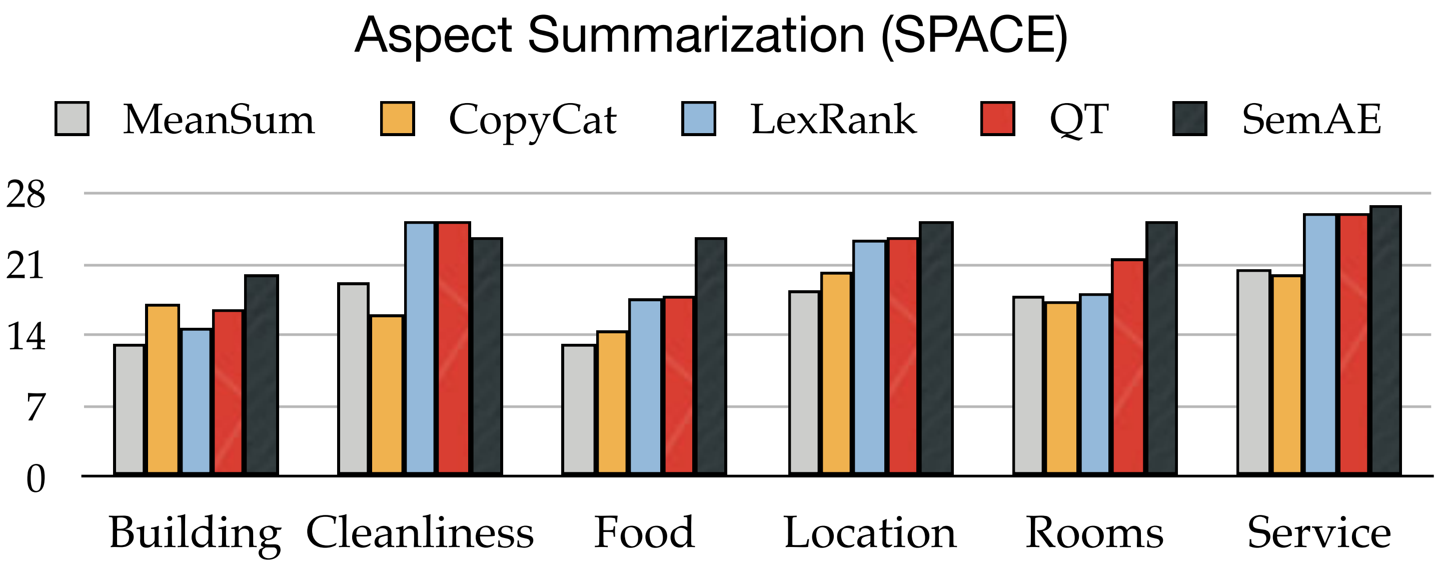
$\mathscr{R}(\alpha^s)$ = [Relevance] - [Redundancy] + [Aspect-awareness]

Aspect Summarization:

$\mathscr{R}(\alpha^s)$ = [Relevance] + [Informativeness]

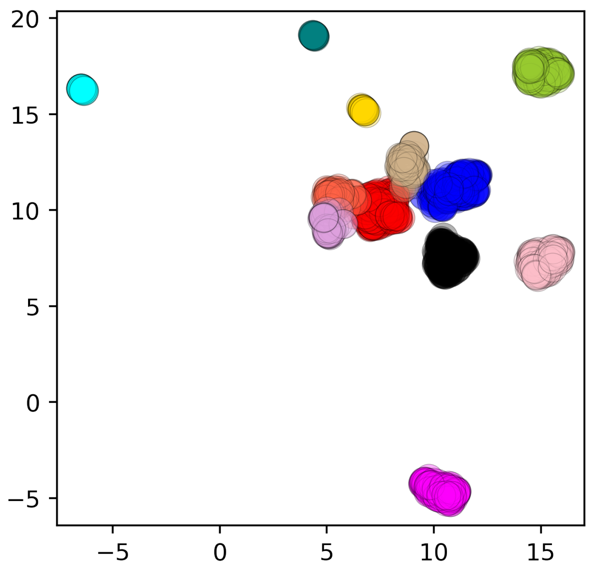
Evaluations





Analysis

- Dictionary representations converges into clusters
- Clusters capture distinct semantic meanings
- Further analysis show that it captures both coarse/fine-grained semantics



Conclusion

- SemAE learns sentence representations as a distribution over latent semantic units
- Sentence selection is performed using information-theoretic metrics
- SemAE achieves strong performance on SPACE and Amazon opinion summarization datasets
- SemAE is able to perform different forms of controllable summarization

ACL 2022 @ Dublin, Ireland