Exploiting Category Specific Information for Guided Summarization

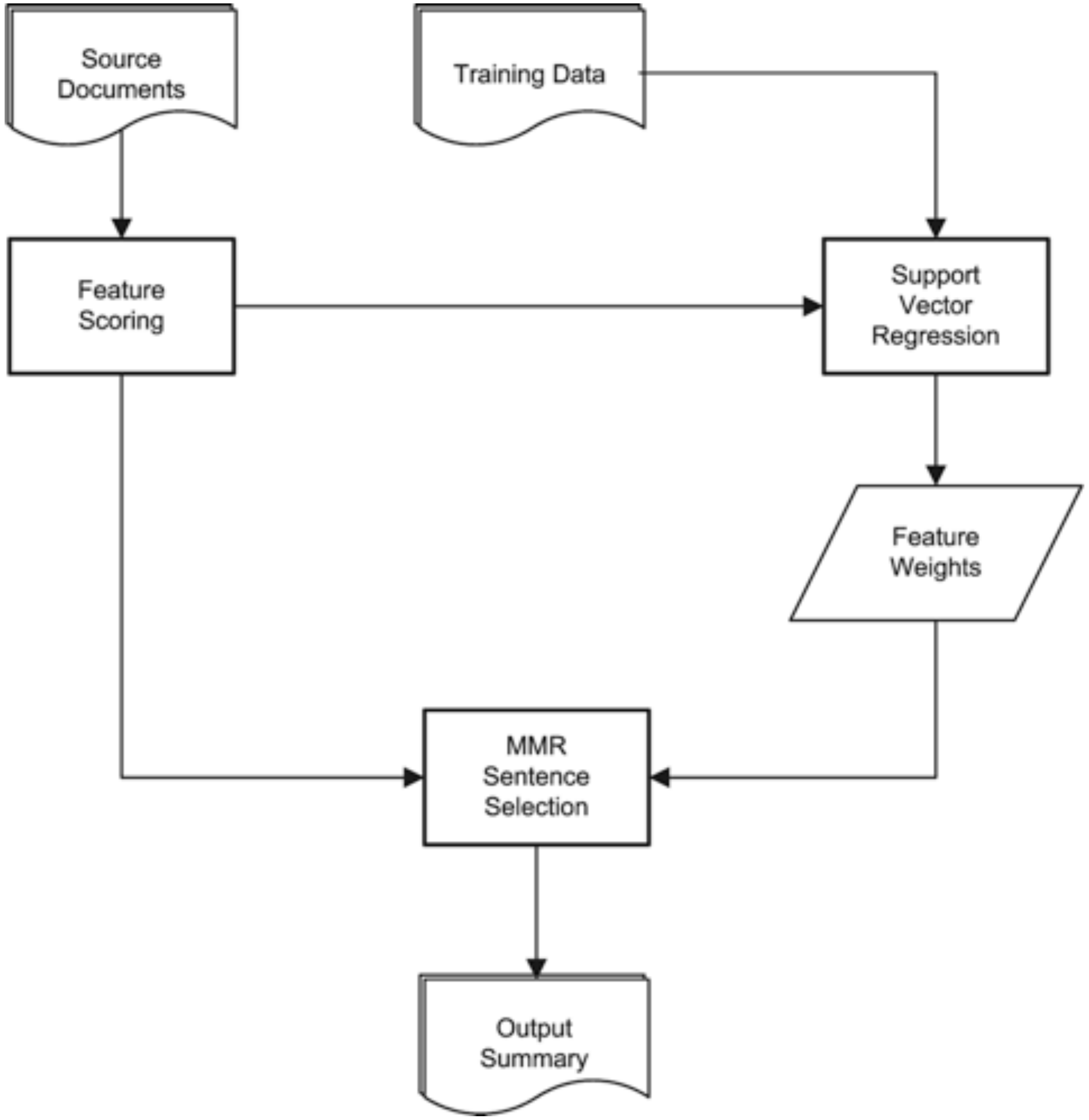
Jun-Ping Ng Praveen Bysani Ziheng Lin Min-Yen Kan Chew-Lim Tan

National University of Singapore

Outline

- System Overview
- Category Specific Features
- Evaluation and Discussion

System Overview



Hypothesis

- Word frequency distribution across different categories should be different
- Some words are more important in certain categories
- e.g. ‘health’ is more salient in “Health and Safety Issues”

What are those words?

Category	Attacks	Health	Endangered
	people	people	years
	minister	food	state
	told	years	national
	government	new	---
	two	health	water

A Hint of Sentence Saliency

- Two ways to look at the difference in word distribution
  - Frequency - Words which are used more are more important
  - Difference in usage - Words which are used differently from the “usual” are more important

Category Specific Information

- Category Relevance Score
- Category KL-Divergence

Category Relevance Score

- Intuition - A word that appears across many documents within a topic and category is more useful
- Linearly weight topic and document frequency scores

$$\frac{\beta(\sum_{w \in s} TFS_c(w)) + (1-\beta)(\sum_{w \in s} CDFS_c(w))}{|s|}$$

Category KL-Divergence

- Intuition - The use of a word varies according to the category an article is written in.
- KL-Divergence between frequency of word across all categories vs specific category

$$CKLD(s) = \sum_{w \in s} p_c(w) log \frac{p_c(w)}{p_C(w)}$$

Generic Features

- Bigram document frequency
  - Backoff model with unigram and bigram document frequencies

$$\frac{\alpha(\sum_{w_u \in s} DFS(w_u)) + (1-\alpha)(\sum_{w_b \in s} DFS(w_b))}{|s|}$$

- Sentence position
- Sentence length

Update Summarization

- Update summaries generated in similar fashion
- But we take into account existing snippets from Set A

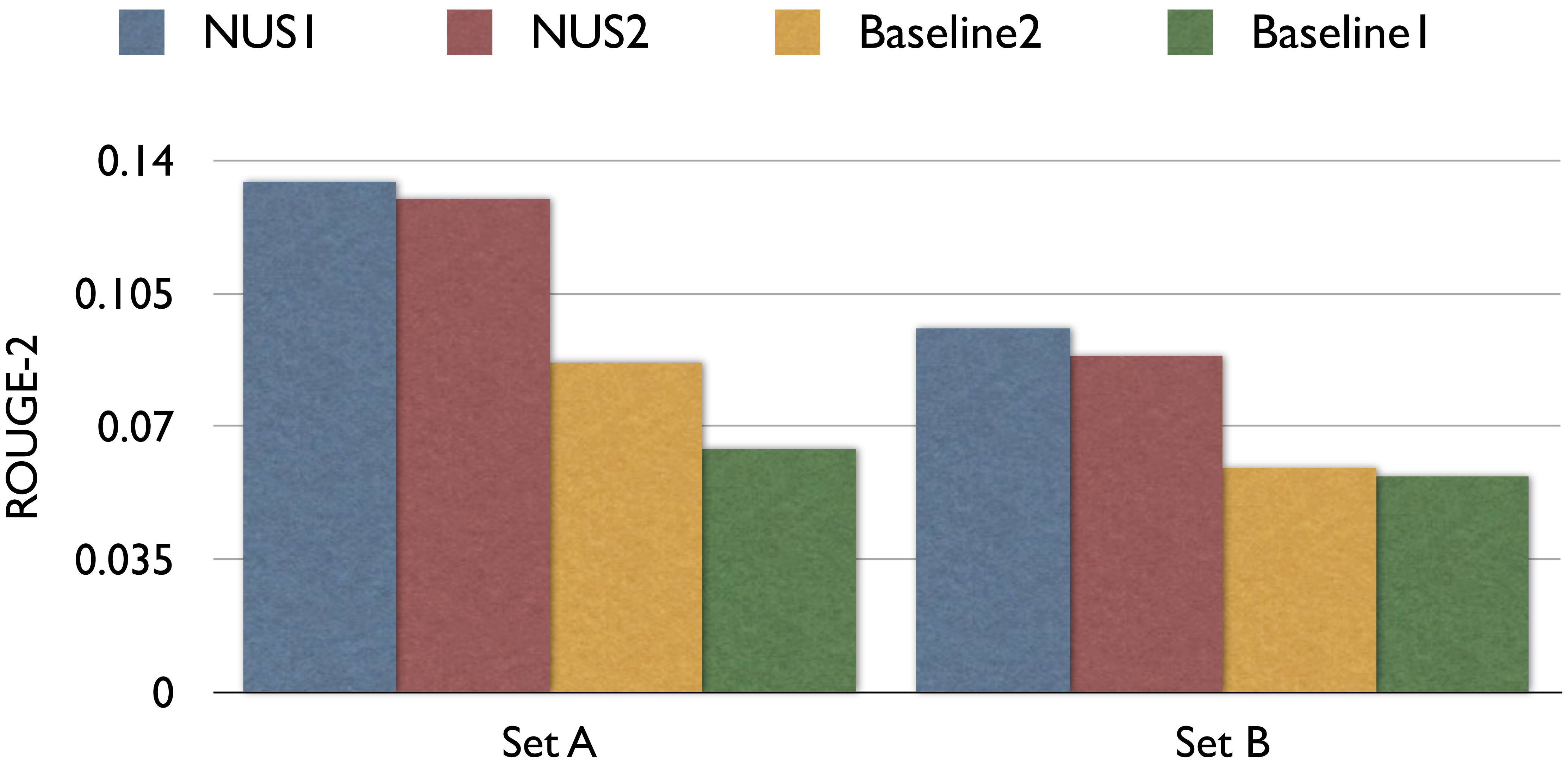
$$\begin{aligned} MMR(s) &= \underline{Score(s) - \lambda \cdot R2(s, S)} \\ &\quad \underline{-\delta \cdot \max_{s' \in A} R2(s, s')} \end{aligned}$$

Typical MMR

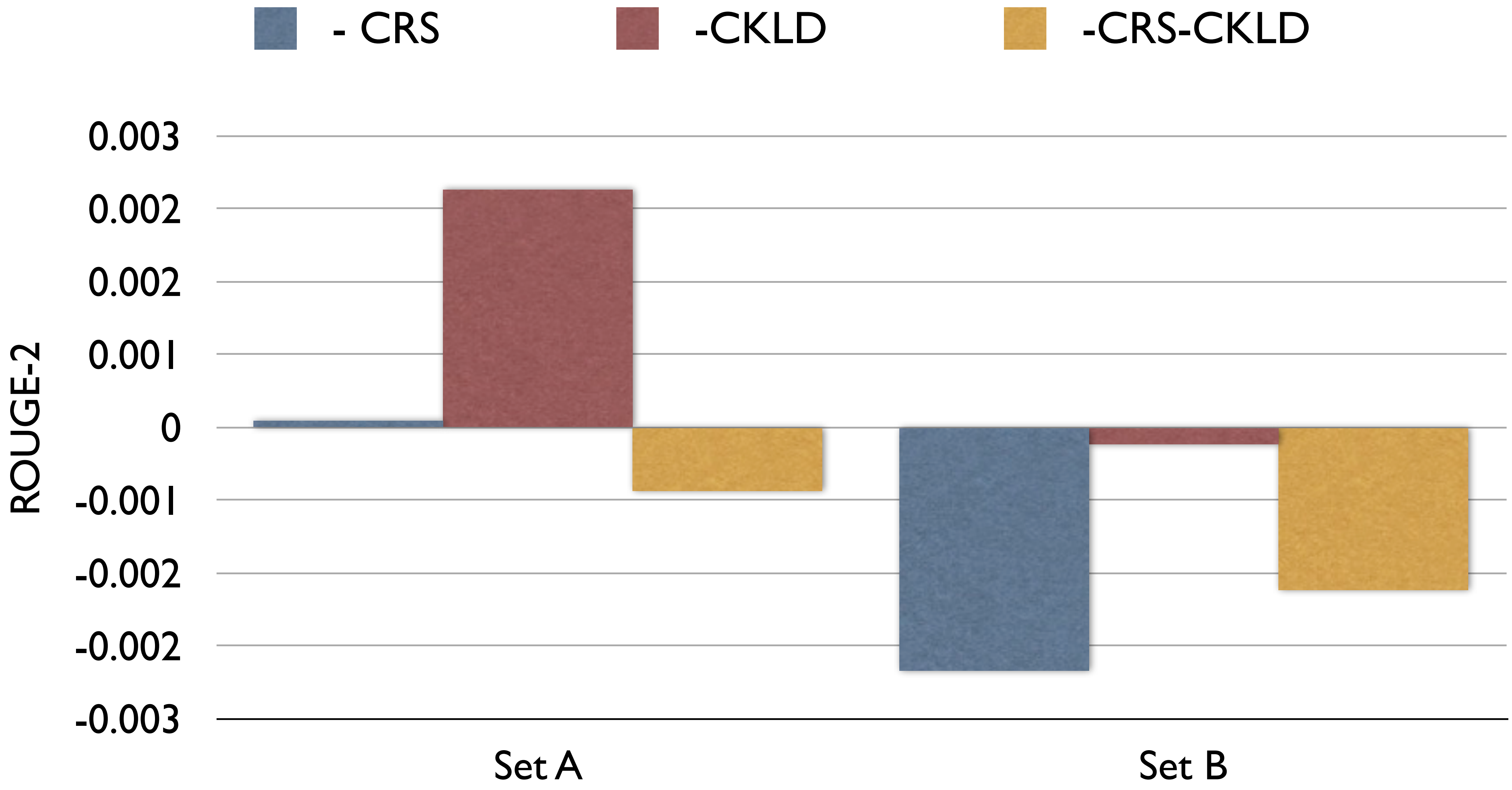
Penalise sentences similar to those in Set A

Evaluation

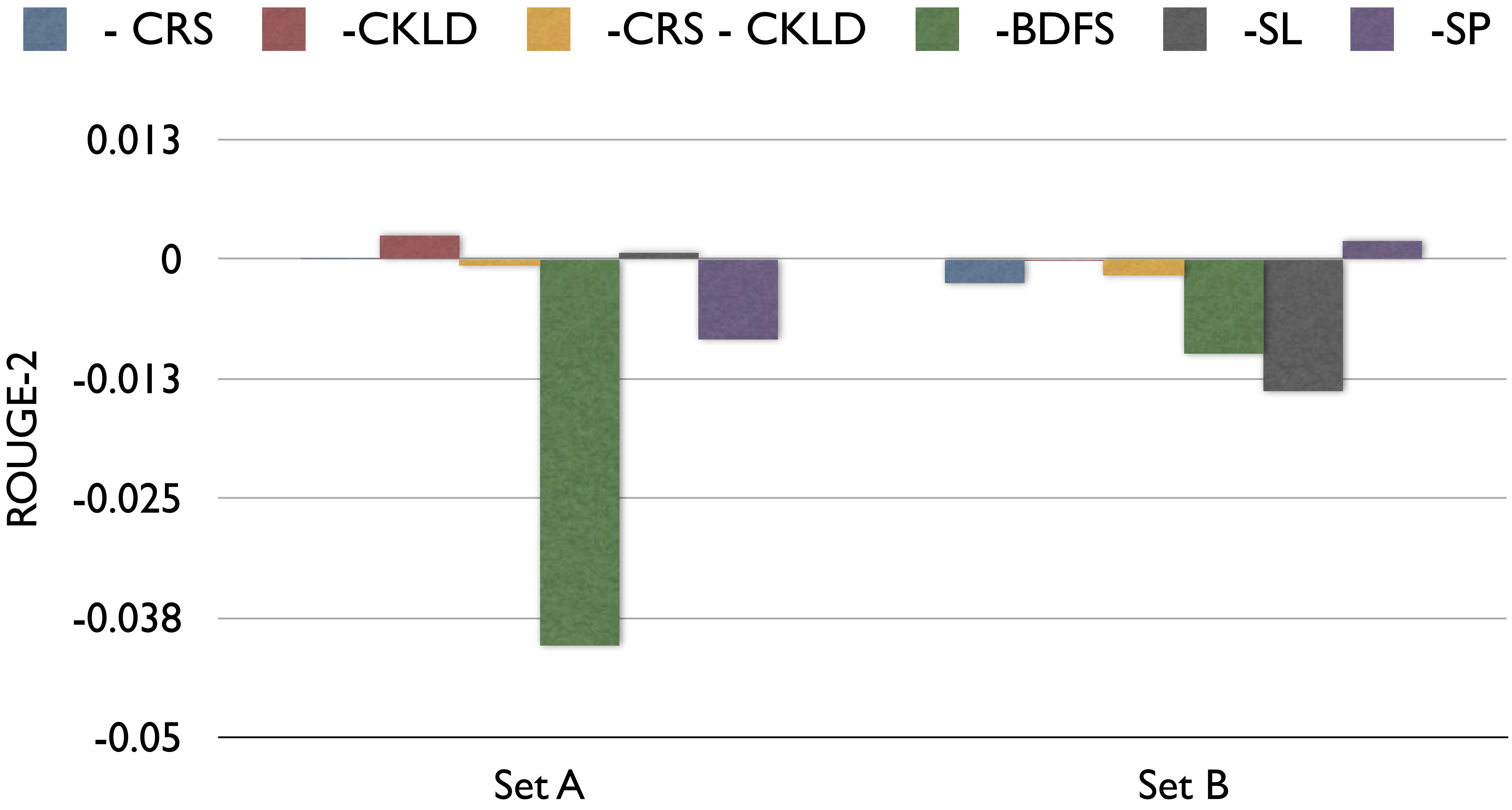
- Against ROUGE-2



What is Important?



All Features



Future Work

- Do better studies to determine influence of category specific information
- Exploit aspect-level information

Thank You

- Word distribution within and outside a category plays a significant role in sentence selection
  - Category relevance score
  - Category KL-Divergence score