Discrete Optimization for Unsupervised Sentence Summarization with Word-Level Extraction

Raphael Schumann[1], Lili Mou[2], Yao Lu[3], Olga Vechtomova[3], Katja Markert[1]

[1]Institute of Computational Linguistics, Heidelberg University
[2]University of Alberta & Alberta Machine Intelligence Institute
[3]University of Waterloo, Canada

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Sentence Summarization

source sentence: the world 's biggest miner bhp billiton announced tuesday it was dropping its controversial hostile bid for rival rio tinto due to the state of the global economy

summary: bhp billiton drops rio tinto takeover bid

- Supervised: source sentences aligned with target summary as training data
- **Unsupervised**: unaligned training data

Sentence Summarization

Abstractive:

- $\text{argmax}_{\mathbf{y}} f(\mathbf{y}|\mathbf{x})$
- big solution space, hard to optimize

Extractive:

- $\text{argmax}_{\mathbf{y}} f(\mathbf{y}{=}(x_i, ..., x_s)|\mathbf{x})$

Word extraction with original order:

source sentence: the world 's biggest miner bhp billiton announced tuesday it was dropping its controversial hostile bid for rival rio tinto due to the state of the global economy

summary: bhp billiton drops rio tinto takeover bid

Approach

- define **search objective** that models quality of a summary
  - language model perplexity
  - semantic similarity
  - hard length constraint
- define **search space**
  - word extraction with original order
- **discrete search** towards optimal solution
  - first choice hill climbing

Search Objective

Summary desiderata:

- **Fluency**: the summary is grammatical
- **Faithfulness**: the summary is similar to the source sentence with respect to meaning
- **Length**: shorter than the source sentence

Search Objective

Model those desiderata:

$$f(\mathbf{y}; \mathbf{x}, s) = f_{\overleftrightarrow{\mathrm{LM}}}(\mathbf{y}) \cdot f_{\mathrm{SIM}}(\mathbf{y}; \mathbf{x})^{\gamma} \cdot f_{\mathrm{LEN}}(\mathbf{y}; s)$$

- $\mathbf{x}$: source sentence
- $\mathbf{y}$: summary sequence
- s: summary length predefined and not subject to optimize

Search Objective: Fluency

Language Model measures grammatical correctness of summary **y**

$$f_{\overleftrightarrow{\mathrm{LM}}}(\mathbf{y}) = \overleftrightarrow{\mathrm{PPL}}(\mathbf{y})^{-1}$$

$$\overleftrightarrow{\mathrm{PPL}}(\mathbf{y}) = \sqrt[2|\mathbf{y}|]{\prod_{i}^{|\mathbf{y}|} \frac{1}{p_{\overrightarrow{\mathrm{LM}}}(y_i|\mathbf{y}_{<i})} \prod_{i}^{|\mathbf{y}|} \frac{1}{p_{\overleftarrow{\mathrm{LM}}}(y_i|\mathbf{y}_{>i})}}$$

- geometric average of forward and backward perplexity
- instantiated by LSTM

Search Objective: Faithfulness

Faithfulness measured as **semantic similarity** between **x** and **y**

$$f_{\mathrm{SIM}}(\mathbf{y}; \mathbf{x}) = \cos(\boldsymbol{e}(\mathbf{x}), \boldsymbol{e}(\mathbf{y}))$$

- cosine similarity between sentence embeddings of **x** and **y**
- instantiated by sent2vec

Search Objective: Length

Hard constraint on the number of words in the summary $\mathbf{y}$

$$f_{\text{LEN}}(\mathbf{y}; s) = \begin{cases} 1, & \text{if } |\mathbf{y}| = s, \\ -\infty, & \text{otherwise.} \end{cases}$$

Search Space

Size of search space for summary length s:

- abstractive: $|\mathcal{V}|^s$, where $\mathcal{V}$ is the vocabulary
- **extractive** with original order: $\binom{n}{s}$
  - summary **y** is a subsequence of sentence **x**

Search Space

- vector $\boldsymbol{a} = (a_1, \ldots, a_n) \in \{0, 1\}^n$ is a Boolean filter over source sentence $\mathbf{x}$
- $\mathbf{y} = \mathbf{x}_{\boldsymbol{a}}$
- enforce length constraint: $\sum_i a_i = s$

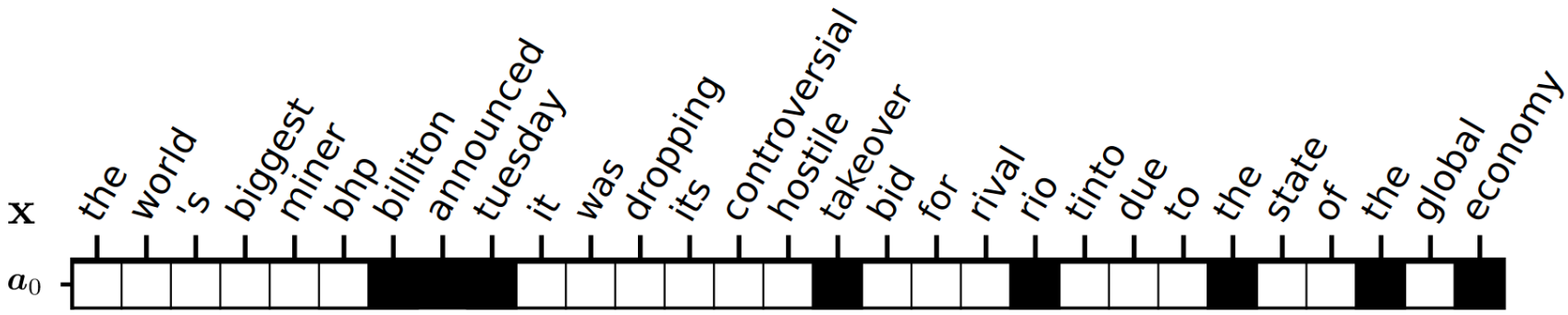
$\boldsymbol{a} = [0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]$

source sentence: the world 's biggest miner bhp billiton announced tuesday it was dropping its controversial hostile bid for rival rio tinto due to the state of the global economy

summary: bhp billiton drops rio tinto takeover bid

Search Algorithm

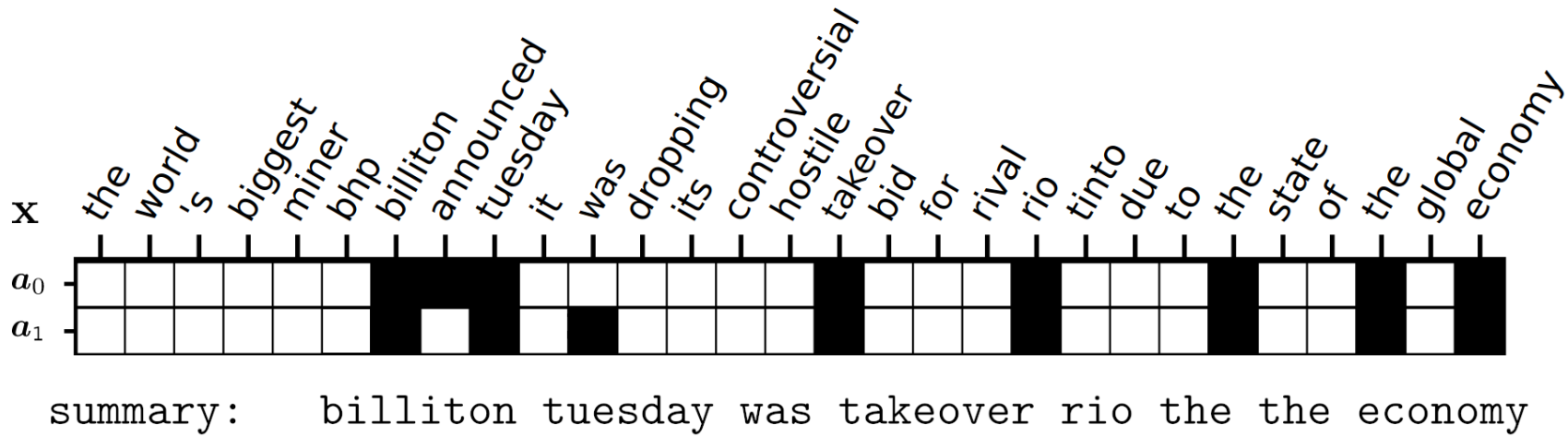
Search towards optimal summary by **first-choice hill climbing**



summary: billiton announced tuesday takeover rio the the economy

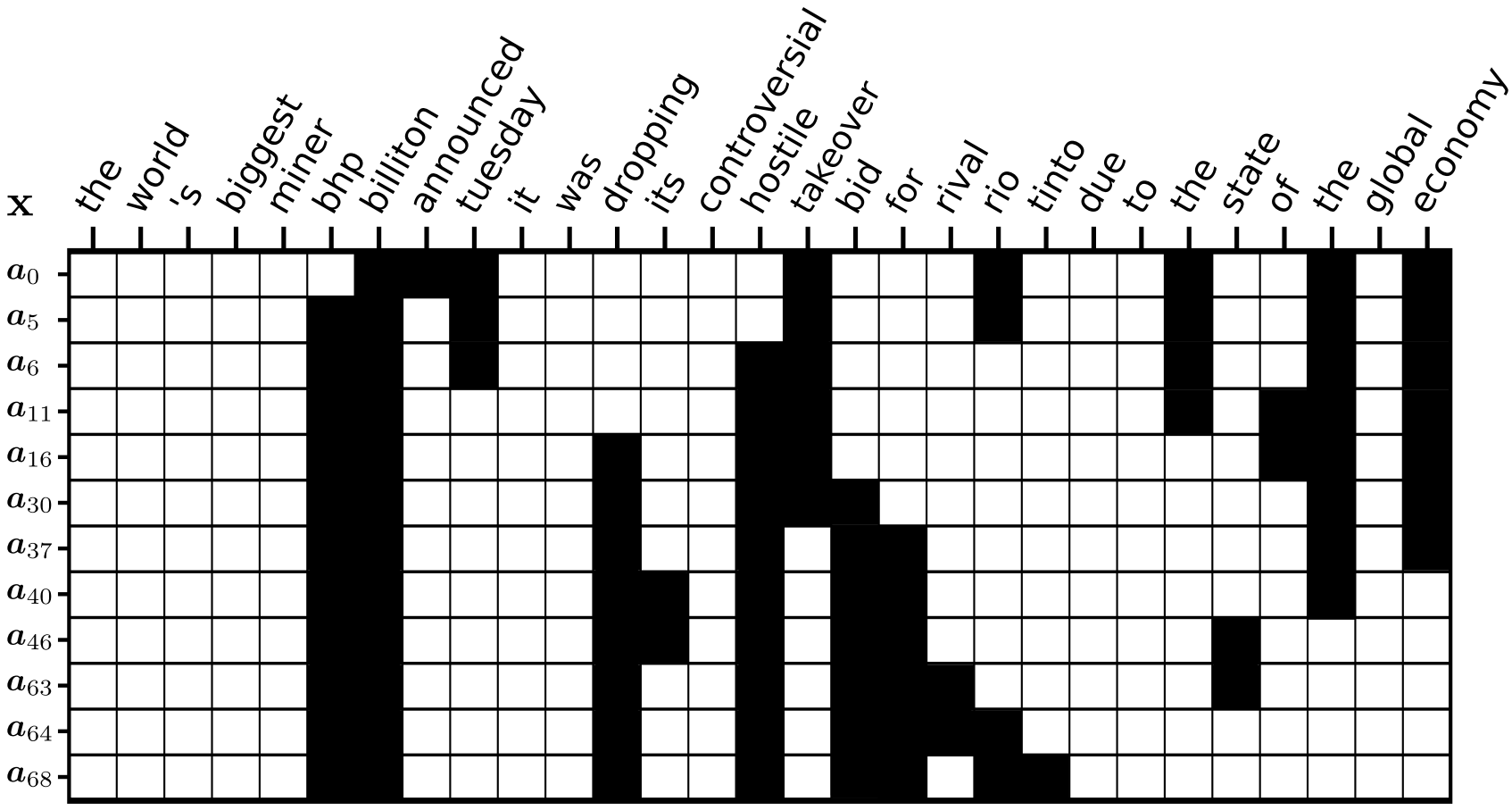
Search Algorithm

Search towards optimal summary by **first-choice hill climbing**



Search Algorithm

Search towards optimal summary by **first-choice hill climbing**



summary: bhp billiton dropping hostile bid for rio tinto

reference: bhp billiton drops rio tinto takeover bid

Evaluation Framework

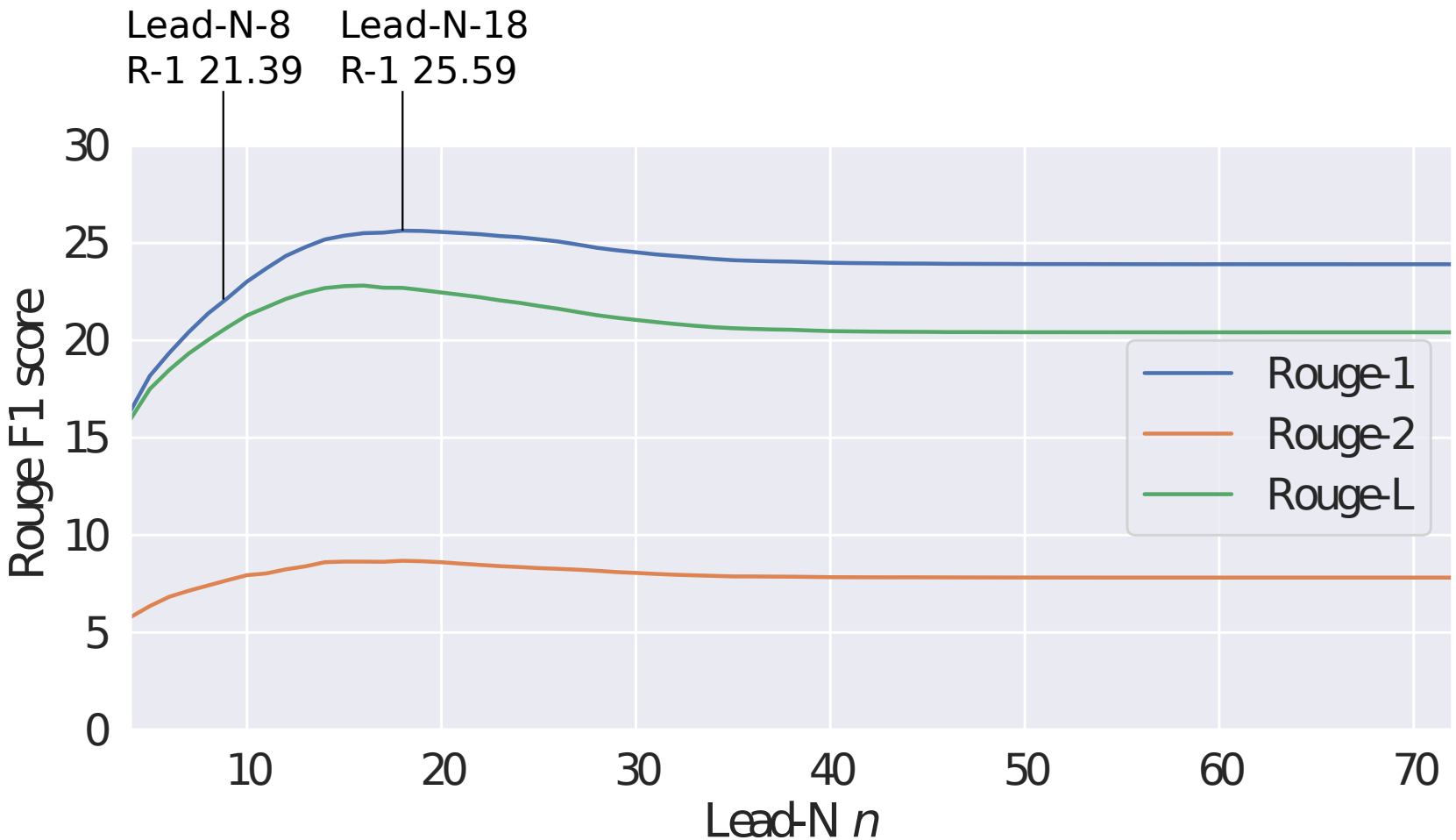
DUC2004:

- sentence paired with 4 human written summaries
- 500 test instances
- Lead-C-75 baseline
- truncated (75 chars) ROUGE Recall scores

Gigaword Headline Generation:

- first sentence of an article paired with its headline
- 3.8M 'train' instances, 1951 test instances
- Lead-N-8 baseline
- ROUGE F1 scores

Evaluation Framework: ROUGE-F1



Evaluation Framework: ROUGE F1

Difference of 4.2 ROUGE-1 F1 points between widely adopted Lead-N-8 baseline and optimal length at $N = 18$

$\Rightarrow$ Need to compare models in context of summary length

Experiment: Our Model

Two model settings:

- *HC_article*
  - language model and sentence embeddings trained on the **source** side (3.8M first article sentences)
- *HC_title*
  - language model and sentence embeddings trained on the **target** side (3.8M headlines)

HC_data_s: Length constrained s according to average summary length of competing models

Experiment: Results Headline Generation

Model	Data		Len D	ROUGE F1			Len O
	source	target		R-1	R-2	R-L	
Lead-N-8	✓		8	21.39	7.42	20.03	7.9
HC_article_8	✓		8	23.09	7.50	21.29	7.9
HC_title_8		✓	8	**26.32**	**9.63**	**24.19**	7.9

Summarization systems with output length around 8 words

Experiment: Results Headline Generation

Model	Data		Len D	ROUGE F1			Len O
	source	target		R-1	R-2	R-L	
Lead-N-10	✓		10	23.03	7.95	21.29	9.8
Wang and Lee 2018	✓	✓	-	27.29	10.01	24.59	10.8
Zhou and Rush 2019		✓	-	26.48	10.05	24.41	9.3
HC_article_10	✓		10	24.44	8.01	22.21	9.8
HC_title_10		✓	10	**27.52**	**10.27**	**24.91**	9.8

Summarization systems with output length around 10 words

Experiment: Results Headline Generation

Model	Data		Len D	ROUGE F1			Len O
	source	target		R-1	R-2	R-L	
Lead-P-50	✓		50%	24.97	**8.65**	22.43	14.6
Fevry and Phang 2018	✓		50%	23.16	5.93	20.11	14.8
Baziotis et al. 2019	✓		50%	24.70	7.97	22.14	15.1
HC_article_50p	✓		50%	**25.58**	8.44	**22.66**	14.9

Summarization systems with output length around 50% of input words

Experiment: Results Headline Generation

Model	Data		Len D	ROUGE F1			Len O
	source	target		R-1	R-2	R-L	
Lead-N-8	✓		8	21.39	7.42	20.03	7.9
Lead-P-50	✓		50%	24.97	**8.65**	22.43	14.6
Fevry and Phang 2018	✓		50%	23.16	5.93	20.11	14.8
Baziotis et al. 2019	✓		50%	24.70	7.97	22.14	15.1
HC_article_50p	✓		50%	**25.58**	8.44	**22.66**	14.9

Comparison to Lead-N-8 baseline can be misleading

Experiment: Results DUC2004

Model	ROUGE Recall		
	R-1	**R-2**	**R-L**
Lead-C-75	22.50	6.49	19.72
SEQ3 (Baziotis et al. 2019)	22.13	6.18	19.3
TOPIARY (Zajic et al., 2004)	25.12	6.46	20.12
BOTTLESUM EX (West et al. 2019)	22.85	5.71	19.87
HC_article_13	24.21	6.63	21.24
HC_title_13	**26.04**	**8.06**	**22.90**

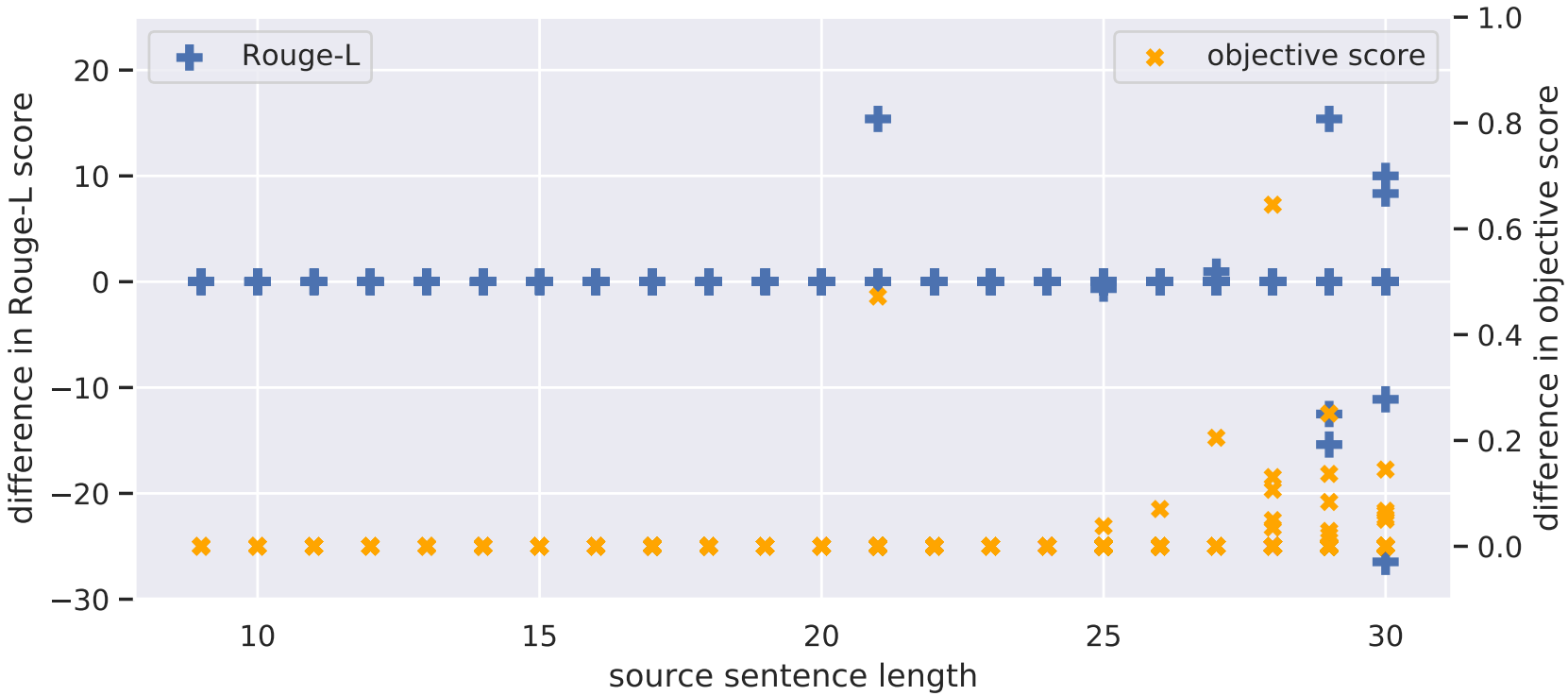
Length 13 words because this is closest to 75 characters

Analysis: Ablation Study

Objective	ROUGE F1 scores		
$f =$	**R-1**	**R-2**	**R-L**
$f_{\overleftrightarrow{\text{LM}}} \cdot f_{\text{SIM}}$ (full model)	**27.52**	**10.27**	**24.91**
$f_{\overleftrightarrow{\text{LM}}}$	25.24	8.87	23.09
f_{SIM}	20.31	4.08	18.19

- *HC_title_10* on the headline generation test set
- Length constraint term omitted from notation.

Analysis: Exhaustive Search



- 1135 instances of Gigaword test set with length ≤ 30
- 5.8M candidates for source sentence length 30 and summary length 8
- 3% of times FCHC doesn't find global optimum

Analysis: Example

Input: a german registered container ship ran aground at the entrance to the french port of le havre early tuesday , but authorities said there were no casualties

Reference: container ship runs aground in french port
HC_article_10: **a** container ship ran aground but **there were** no casualties
HC_title_10: container ship ran aground **at french port** but no casualties
HC_title_8: ship ran aground at french port no casualties

Takeaways

- state-of-the-art for unsupervised sentence summarization in terms of ROUGE scores
- word extraction with original order is a feasible strategy
- knowing the target distribution is a big advantage
- ROUGE F1 evaluation only in context of summary length