A Simple Theoretical Model of Importance for Summarization

Maxime Peyrard

ACL19 Outstanding Paper

Xiachong Feng

Author

Maxime Peyrard
EPFL
洛桑联邦理工学院

2019

pdf bib abs **A Simple Theoretical Model of Importance for Summarization**
Maxime Peyrard
Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

pdf bib abs **Studying Summarization Evaluation Metrics in the Appropriate Scoring Range**
Maxime Peyrard
Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

pdf bib abs **MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance**
Wei Zhao | Maxime Peyrard | Fei Liu | Yang Gao | Christian M. Meyer | Steffen Eger
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)

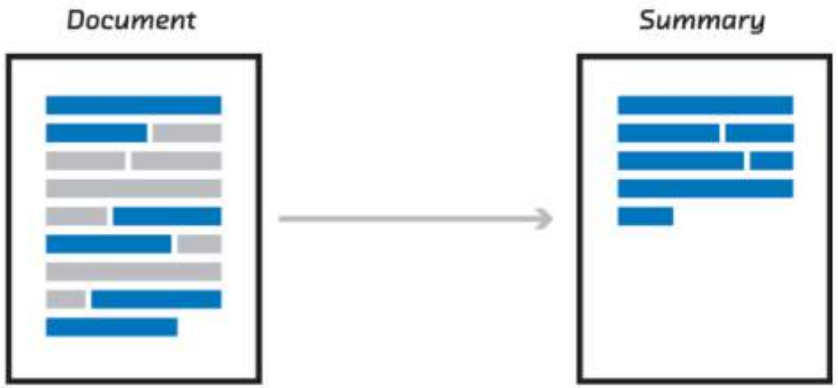
2018

pdf bib **Live Blog Corpus for Summarization**
Avinesh P.V.S. | Maxime Peyrard | Christian M. Meyer
Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)

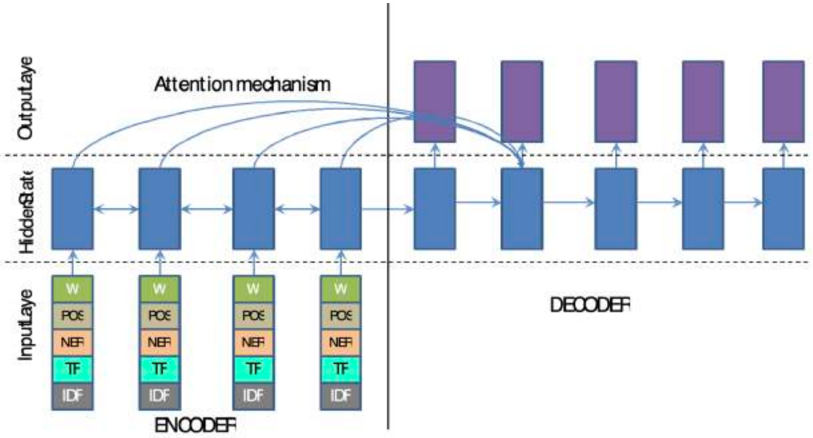
pdf bib abs **Objective Function Learning to Match Human Judgements for Optimization-Based Summarization**
Maxime Peyrard | Iryna Gurevych
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)

Overview

Extractive



Abstractive



ROUGE

Document

Summary

BLEU

Document

Summary

Importance

Document

Summary

Summarization

Summarization is the process of **identifying the most important information** from a source to **produce a comprehensive output** for a particular user and task.

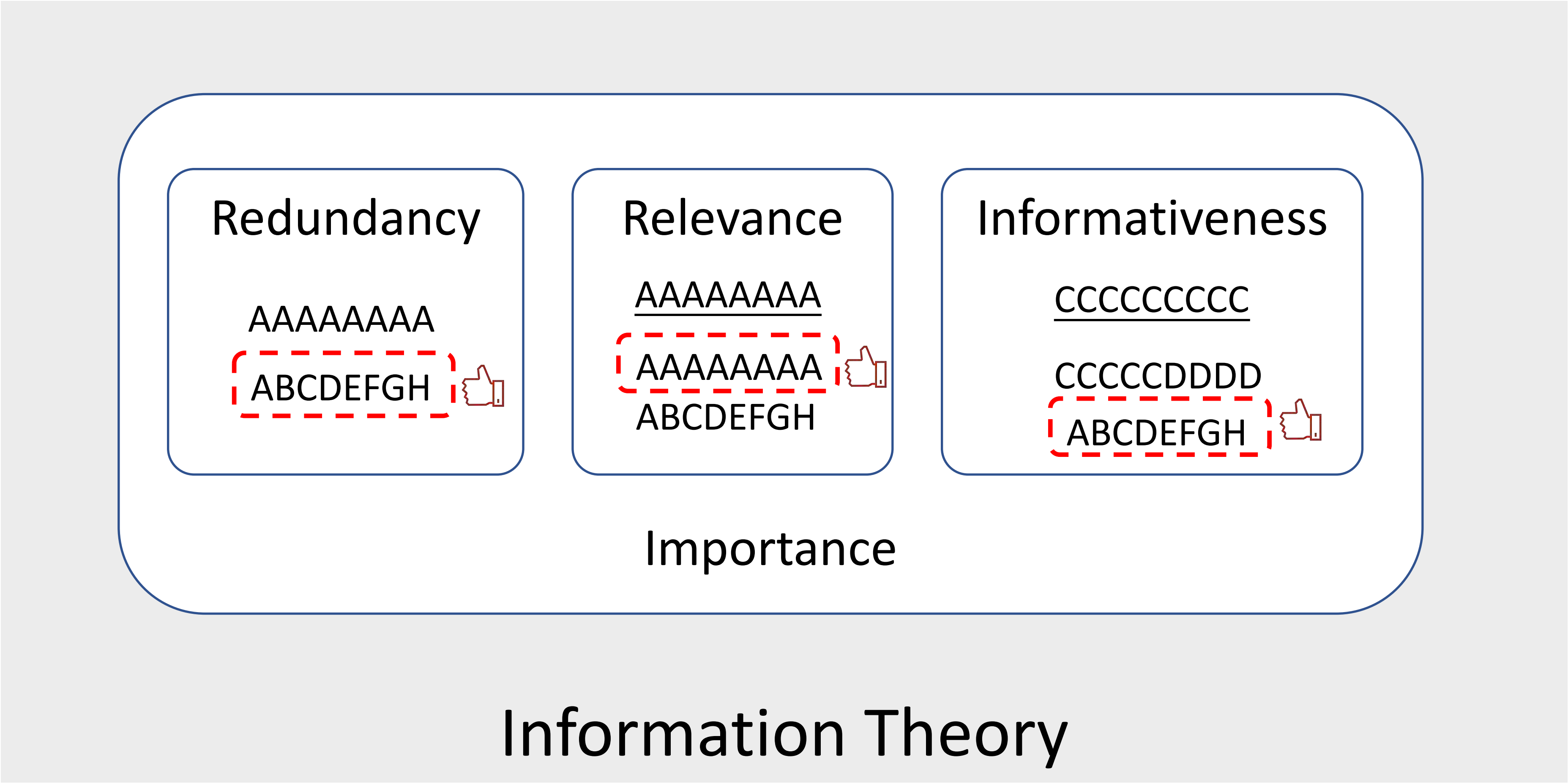
Summarization

Summarization is the process of **identifying the most important information** from a source to **produce a comprehensive output** for a particular user and task.

The core challenge of summarization

Natural Language Generation

Overview


Redundancy
AAAAAAAA
ABCDEFGH
Relevance
AAAAAAAA
AAAAAAAA
ABCDEFGH
Informativeness
CCCCCCCCC
CCCCCDDDD
ABCDEFGH
Importance
Information Theory

Information theory

- Entropy for event

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i))$$

e.g. $X =$ 抛一枚硬币

$x_1 =$ 正面朝上

$x_2 =$ 反面朝上

$$\overbrace{x_1 \qquad\qquad x_2}^{X}$$

- Entropy for text $X = w_1, w_2, \ldots, w_n$

$$\overbrace{w_1 \quad w_2 \quad \cdots\cdots \quad w_n}^{X}$$

$$p(X) = p(w_1)p(w_2)\cdots p(w_n)$$

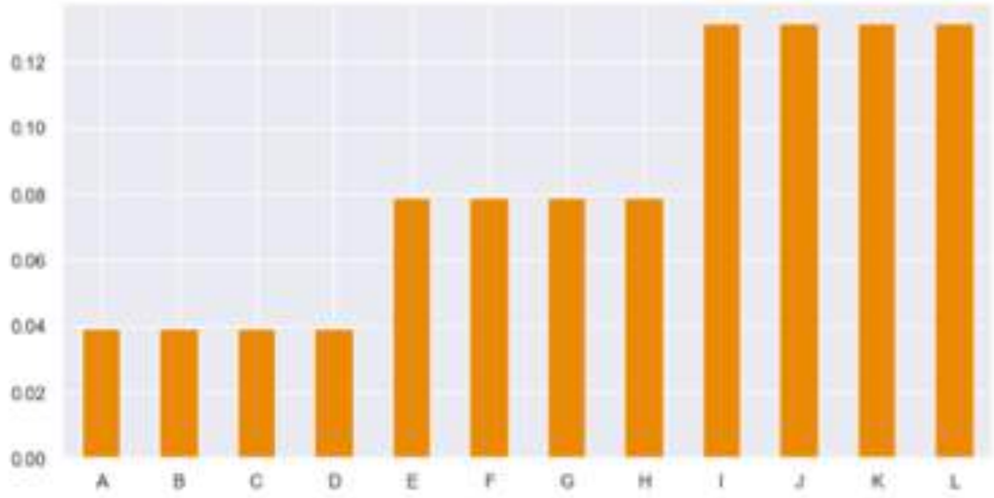
$$H(X) = -\sum_{i=1}^{n} p(\boxed{w_i})\log(p(\boxed{w_i}))$$

Semantic unit

Semantic Units Ω

- Atomic piece of information Ω

- Words
- Characters
- BPE
- Topic models
- Frame semantics
- ……



$$H(X) = -\sum_{i=1}^{n} p(\omega_i)\log(p(\omega_i))$$

Semantic unit

- X can be represented by a probability distribution $\mathbb{P}_X$ over the semantic units Ω.

Notation

- Semantic Unit $\omega_i \in \Omega$
- Source document(s) D, $\mathbb{P}_D$
- Candidate summary S, $\mathbb{P}_S$

Redundancy

Redundancy

AAAAAAAA

ABCDEFGH

- A summary should contain a lot of information.
- For a summary S represented by $\mathbb{P}_S$:

$$H(S) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i)\log(\mathbb{P}_S(\omega_i))$$

- Redundancy

$$Red(S) = -\mathrm{H}(S)$$

$\mathrm{H}(S)$ $\uparrow$ $Red(S)$ $\downarrow$

Redundancy in Previous Works

- Maximum coverage
- MMR (Maximal marginal relevance)
  - The selected sentence is the most important one amongst the remaining sentences and it has the **least content overlap** with the current summary.
- Submodular functions
  - Reward diversity. Reward a higher score when picking a sentence that is not too similar to the summary set.

Relevance

Relevance

AAAAAAAA

AAAAAAAA

ABCDEFGH

- Intuitively, observing a summary should reduce our uncertainty about the original text.

$$Rel(S,D) = -CE(S,D)$$

$$Rel(S,D) = \sum_{\omega_i} \mathbb{P}_S(\omega_i)\log(\mathbb{P}_D(\omega_i))$$

$CE(S,D) \downarrow$ $Rel(S,D) \uparrow$

Informativeness

- Intuitively, a summary is informative if it induces, for a user, a great change in her knowledge about the world.
- K the background knowledge $\mathbb{P}_K$

$$Inf(S,K) = CE(S,K)$$

$$Inf(S,K) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i)\log(\mathbb{P}_K(\omega_i))$$

Informativeness

CCCCCCCCC

CCCCCDDDD

$\boxed{\text{ABCDEFGH}}$

Importance

$$Red(S) = -\mathrm{H}(\mathrm{S})$$

$$Rel(S, D) = -CE(S, D)$$

$$Inf(S, K) = CE(S, K)$$

Importance

$$D \qquad K$$

Source Document
Background knowledge

$$\Omega = \omega_1 \ \omega_2, \cdots, \ \omega_n$$

Semantic Units

$$\mathbb{P}_D \qquad \mathbb{P}_K$$

Distribution

$$d_i = \mathbb{P}_D(\omega_i) \qquad k_i = \mathbb{P}_K(\omega_i)$$

For one unit ω_i

$$\boxed{f(d_i, k_i)}$$

Importance of unit ω_i

$$f(d_i, k_i)$$

$$d_i = d_j \quad k_i > k_j$$

$$\downarrow$$

$$f(d_i, k_i) < f(d_j, k_j)$$

Informativeness

$$k_i = k_j \quad d_i > d_j$$

$$\downarrow$$

$$f(d_i, k_i) > f(d_j, k_j)$$

Relevance

$$I\big(f(d_i, k_i)\big) = \alpha I(d_i) + \beta I(k_i)$$

Additivity

$$\sum_i f(d_i, k_i) = 1$$

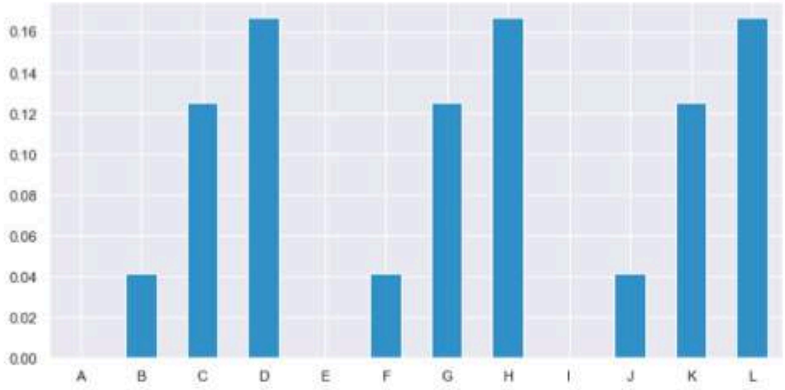
Normalization

$f(d_i, k_i)$

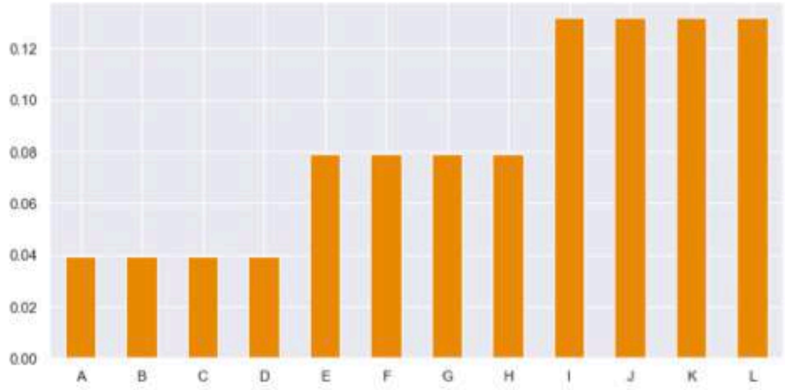
$$\mathbb{P}_{\frac{D}{K}}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^{\alpha}}{k_i^{\beta}}$$

$$C = \sum_i \frac{d_i^{\alpha}}{k_i^{\beta}}, \alpha, \beta \in \mathbb{R}^{+}$$

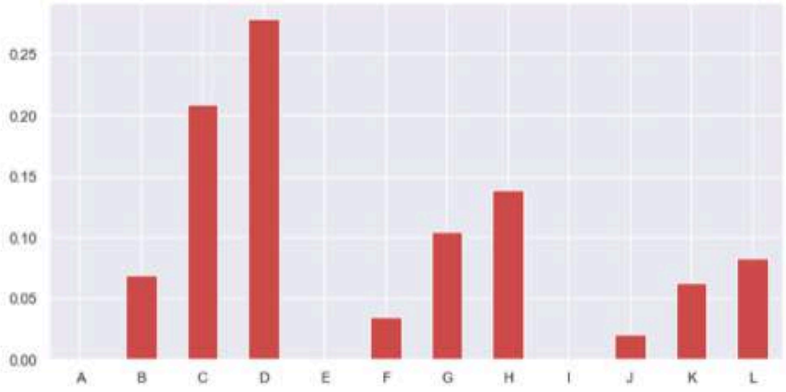
$$\mathbb{P}_{\frac{D}{K}}$$



(a) ditribution $\mathbb{P}_D$



(b) distribution $\mathbb{P}_K$



(c) distribution $\mathbb{P}_{\frac{D}{K}}$

Summary scoring function

$$S \longrightarrow \mathbb{P}_{\frac{D}{K}}$$

$$Red(S) = -\mathrm{H}(\mathrm{S})$$

$$\boxed{\theta_I(S, D, K)} = -KL\left(\mathbb{P}_S \parallel \mathbb{P}_{\frac{D}{K}}\right) = -CE\left(\mathbb{P}_S \parallel \mathbb{P}_{\frac{D}{K}}\right) + H(S)$$

$$S^* = \underset{S}{\operatorname{argmax}}\, \theta_I = \underset{S}{\operatorname{argmin}}\, KL(\mathbb{P}_S \parallel \mathbb{P}_{\frac{D}{K}})$$

Experiments

- TAC-2008 and TAC-2009

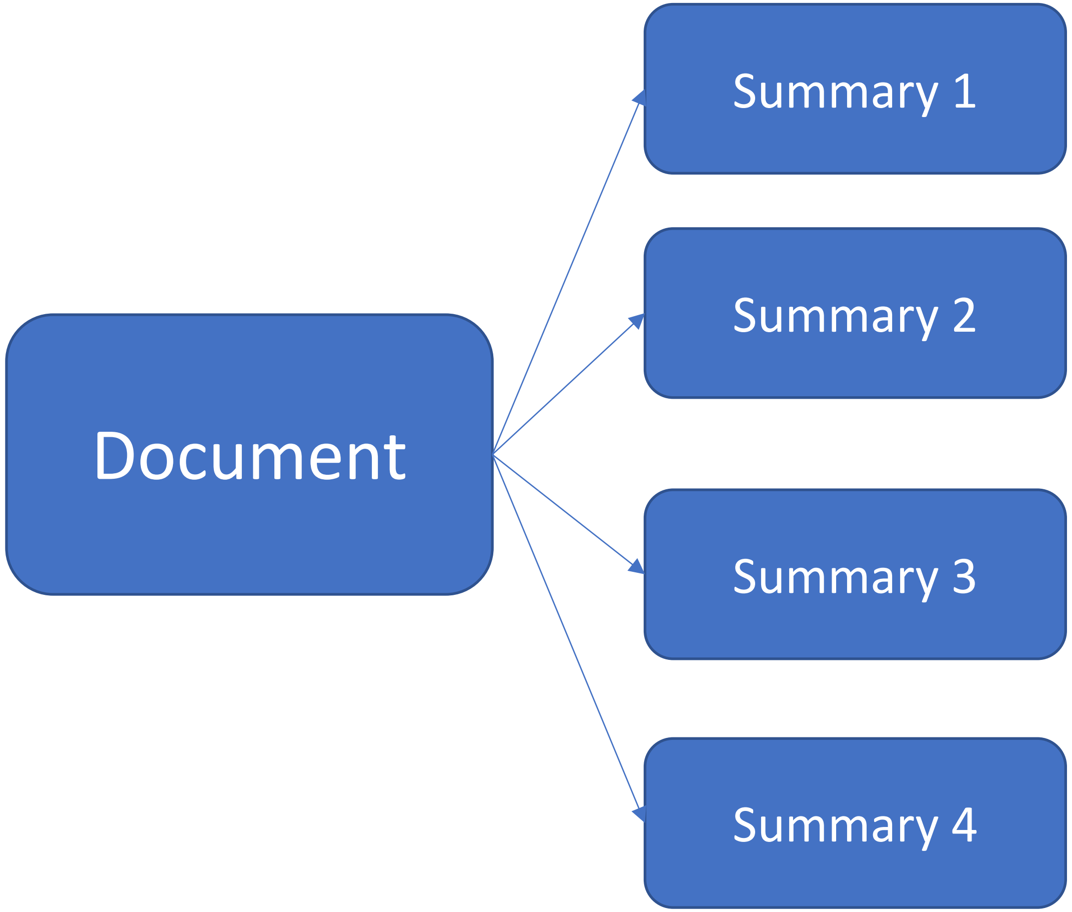
- Generic multi-document summarization
    - A documents (10 documents) --> Summary

- Update multi-document summarization
    - Given A documents (10 documents)
    - B documents (10 documents) --> Summary

Setup and Assumptions

- semantic units : words
- For update summarization, K is the frequency distribution over words in the background documents (A).
- For generic summarization, K is the uniform probability distribution
- $\alpha = \beta = 1$

Correlation with humans



Manual Pyramid scores	θ_I
2	2
1	1
3	4
4	3

Kendall τ

Result

	Generic	Update
ICSI	.178	.139
Edm.	.215	.205
LexRank	.201	.164
KL	.204	.176
JS	.225	.189
KL_{back}	.110	.167
JS_{back}	.066	.187
Red	.098	.096
Rel	.212	.192
Inf	.091	.086
θ_I	**.294**	**.211**

Example

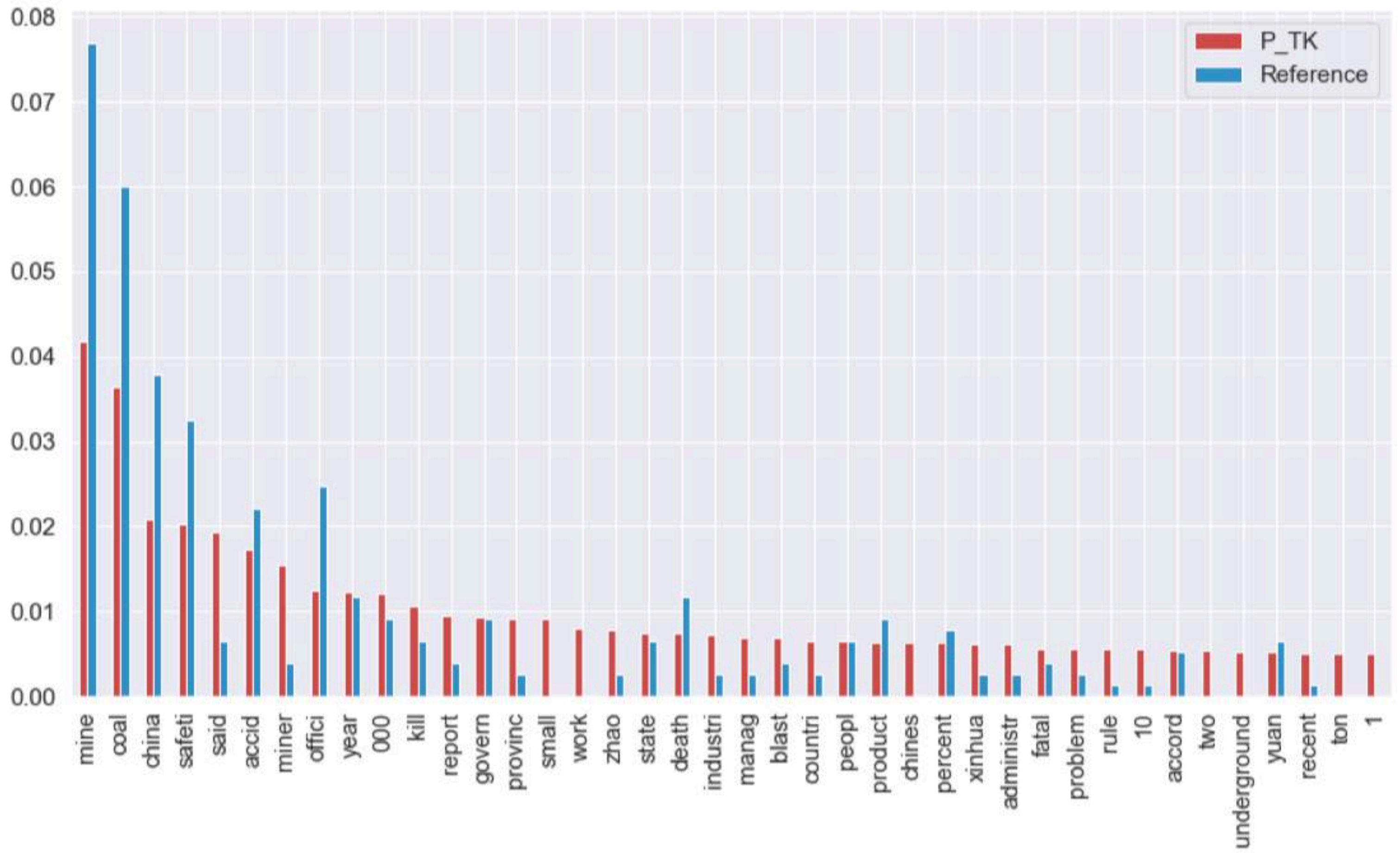


Figure 2: Example of $\mathbb{P}_{\frac{D}{K}}$ in comparison to the word distribution of reference summaries for one topic of TAC-2008 (D0803).

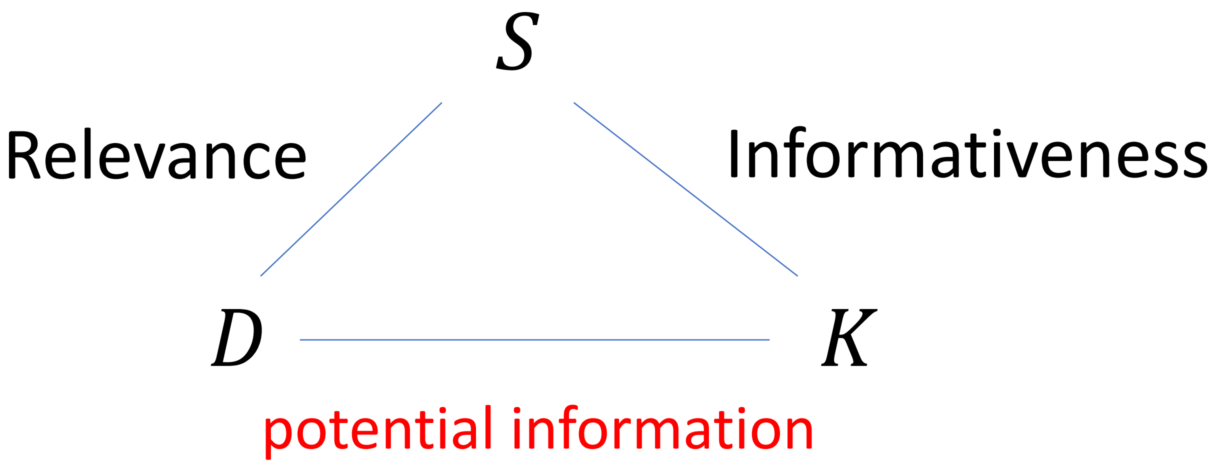
$H(\mathbb{P}_{\frac{D}{K}})$

- Measures the number of possibly good summaries.

- Low : little uncertainty about which semantic units to extract (few possible good summaries).

- High : many equivalently good summaries are possible

Potential Information



$$PI(D,K) = CE(D,K)$$

Thanks!