Which Information Matters? Dissecting Human-written Multi-document Summaries with Partial Information Decomposition

Laura Mascarell Yan L'Homme Majed El Helou

ETH Zurich

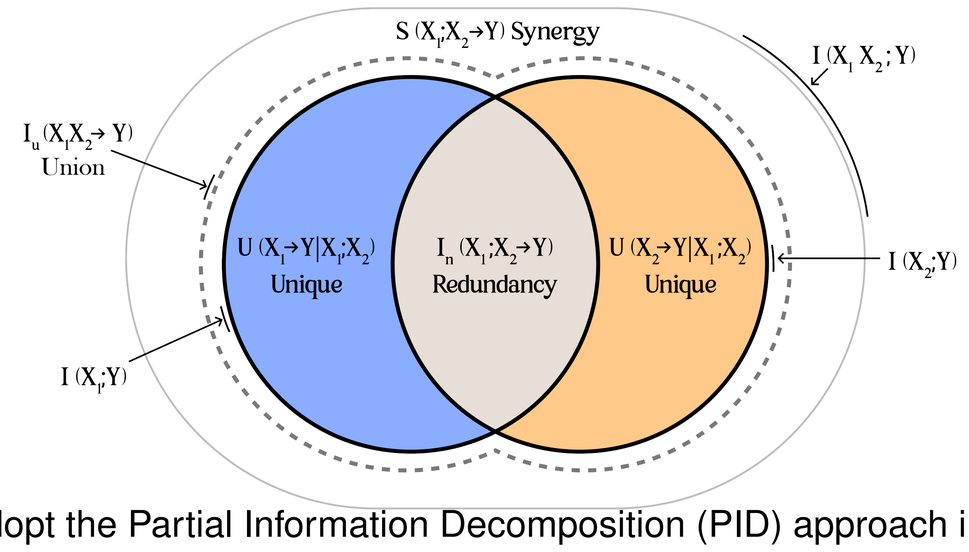
What information constitutes a high-quality multi-document summary?

We propose to categorise the information in a summary into:

- Information provided by a **unique** source,
- Information provided by at least one source (**union**),
- **Redundant** information from all source documents,
- New information derived from considering sources jointly (**synergy**).

Source X_1	Kimchi is fermented cabbage.
Source X_2	Fermented foods are rich in probiotics.
Target Y	Fermented foods, e.g. Kimchi, are rich in probiotics.
Unique X_1	The nature and preparation of kimchi.
Unique X_2	A general characteristic of fermented foods.
Redundancy	Information about *fermented*.
Synergy	Inferring that Kimchi is a fermented food.

Partial Information Decomposition in MDS



We adopt the Partial Information Decomposition (PID) approach in [1] to MDS, considering sentences as units of information.

Let $\mathcal{X} = \{D_1, ..., D_n\}$ be a set of n source documents, where each document is a collection of sentences $D_i = \{d_i^1, ..., d_i^{|D_i|}\}$, and a multi-document summary of m sentences $S = \{s^1, ..., s^m\}$:

Redundancy is the maximum information we can obtain about the summary that is less informative than any of the sources D_i

$$I_{\cap}^{\text{MDS}}(\mathcal{X} \to S) := \sup_{D \in \mathcal{D}} I(S; D) | \forall i,\, D \sqsubset D_i \tag{1}$$

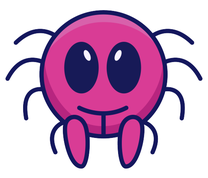
Union is the minimum information we can obtain about the summary that is more informative than any of the sources D_i

$$I_{\cup}^{\text{MDS}}(\mathcal{X} \to S) := \inf_{D \in \mathcal{D}} I(S; D) | \forall i,\, D_i \sqsubset D \tag{2}$$

Unique information U^{MDS} and **Synergy** S^{MDS} are defined using Eq. (1) and Eq. (2), respectively, as in [1]:

$$U^{\text{MDS}}(D_i \to S | \mathcal{X}) = I(S; D_i) - I_{\cap}^{\text{MDS}}, \tag{3}$$

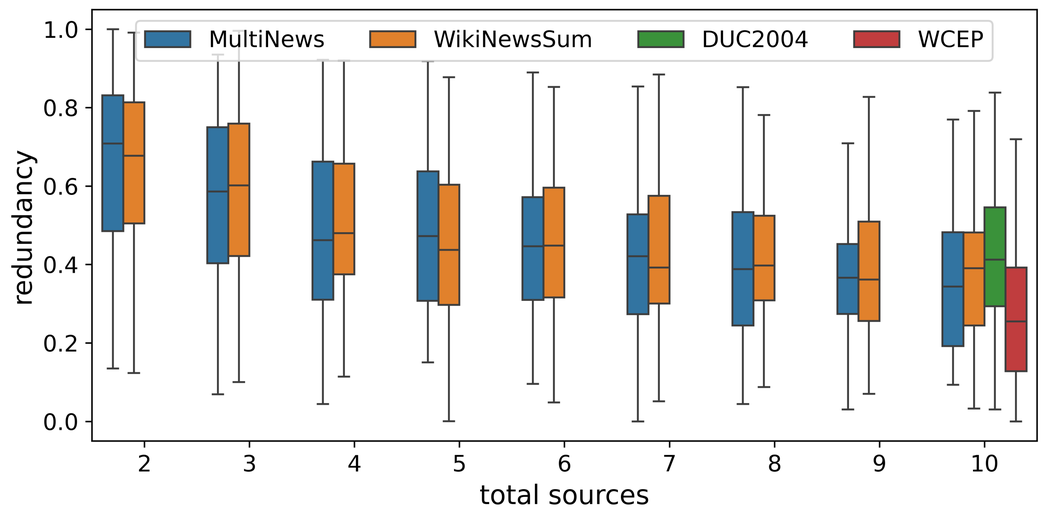
$$S^{\text{MDS}}(\mathcal{X} \to S) = I(S; \mathcal{X}) - I_{\cup}^{\text{MDS}} \tag{4}$$



GitHub: github.com/mediatechnologycenter/SPIDer

This project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, and the ETH Zurich Foundation.

PID of Human-written Summaries



Dataset	Union	Synergy	Redund.	Unique
MultiNews	1.0 (±0.0)	0.0 (±0.0)	0.48 (±0.2)	0.30 (±0.1)
WikiSum	1.0 (±0.0)	0.0 (±0.0)	0.48 (±0.2)	0.30 (±0.1)
DUC2004	1.0 (±0.0)	0.0 (±0.0)	0.43 (±0.2)	0.35 (±0.1)
WCEP	1.0 (±0.0)	0.0 (±0.0)	0.29 (±0.2)	0.41 (±0.2)

→ WCEP is extended with additional source articles not considered in the summaries. Therefore, the scores differ from real MDS datasets.

→ Synergy is negligible, so union represents the total mutual information

→ Redundancy decreases with the number of sources.

→ The more sources, the more they contribute individually (*see paper*).

→ The first three sources contribute the most in all datasets, regardless of the number of source (*see paper*).

Measuring Synergistic Information

MultiRC is a reading comprehension dataset that requires multiple source sentences to identify correct answers (Synergy).

We transform MultiRC into a MDS dataset:

- Source documents: each sentence required to answer a question.
- Summary: a concatenated question-answer pair.

$\mathbf{D}_1$	The story revolves around **A.C.P. Ramakant Chaudhary** whose eldest son Vikas is killed in a pre-planned accident.
$\mathbf{D}_2$	Vishal confronts Baba Khan and the ganglords threaten to eliminate the **A.C.P.** as well as **his wife Revati**.
$\mathbf{S}_{correct}$	Who is Revati's husband? Ramakant Chaudhary
$\mathbf{S}_{incorrect}$	Who is Revati's husband? Baba Khan
$\mathbf{S}_{unrelated}$	How many people comfort the baby? 2

	Union	Synergy	Redund.	Unique
$S_{correct}$	0.05 (±0.3)	0.28 (±0.5)	0.02 (±0.2)	0.04 (±0.1)
$S_{incorrect}$	0.05 (±0.3)	0.26 (±0.5)	0.02 (±0.2)	0.03 (±0.1)
$S_{unrelated}$	0.04 (±0.3)	0.24 (±0.5)	0.01 (±0.2)	0.03 (±0.1)

→ Synergistic information is the dominant information component and correct answers achieve the highest synergy.

References

[1] Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 2022.