Neurocomputing 548 (2023) 126356

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Align-then-abstract representation learning for low-resource summarization

Gianluca Moro *, Luca Ragazzi

Department of Computer Science and Engineering (DISI), University of Bologna, Via dell'Universitá 50, I-47522 Cesena, Italy

ARTICLE INFO

Article history:
Received 3 December 2022
Revised 16 March 2023
Accepted 14 May 2023
Available online 2 June 2023

Keywords:
Long document summarization
Abstractive summarization
Low-resource
Representation learning
NLP

ABSTRACT

Generative transformer-based models have achieved state-of-the-art performance in text summarization. Nevertheless, they still struggle in real-world scenarios with long documents when trained in low-resource settings of a few dozen labeled training instances, namely in low-resource summarization (LRS). This paper bridges the gap by addressing two key research challenges when summarizing long documents, i.e., long-input processing and document representation, in one coherent model trained for LRS. Specifically, our novel align-then-abstract representation learning model (ATHENA) jointly trains a segmenter and a summarizer by maximizing the alignment between the chunk-target pairs in output from the text segmentation. Extensive experiments reveal that ATHENA outperforms the current state-of-the-art approaches in LRS on multiple long document summarization datasets from different domains.

1. Introduction

Abstractive summarization is a natural language processing task of generating a short version of a document while preserving the salient details [1]. The training of modern transformer-based solutions [2–5] typically requires massive labeled data, which levies obstacles in realistic scenarios of *low-resource summarization* (LRS) distinguished by just a few dozen labeled training instances [6,7,43]. Of a particular challenge, LRS in the real world regards long documents [8,9] because producing the ground-truth summary of lengthy texts is expensive, time-consuming, and may demand domain-knowledge proficients. *Long-input processing* is a crux challenge in summarizing extended articles and assumes finding a strategy to address the quadratic memory complexity in the input size of transformer-based models [10]. Existing methods generally rely on input truncation [11,12] or salient content selection [8]. Nevertheless, the reliance on the input given to the model can undesirably prevent proper *document representation learning*, the quality of which may, in turn, affects the summarization accuracy, especially in low-resource conditions. To date, few efforts have been made to resolve this issue.

This work bridges the gap by reconciling the inherent tension between the two highly dependent problems, i.e., long-input processing and representation learning, in one coherent and synergistic model trained for LRS. Our model, *align-then-abstract representation learning* (ATHENA),[1] jointly trains a text segmentation module with an abstractive summarizer with a novel alignment loss. Notably, we (i) address the computational issues of summarizing lengthy documents by segmenting them into small content-wise chunks, which are synthesized with *fewer memory requirements*, (ii) give summarization models entire inputs *without truncating any information* a priori or selecting a subset of sentences, and (iii) *cope with the scarcity of labeled instances* by implicitly augmenting the training data through text segmentation.

To reckon the effectiveness and generality of our solution, we accomplish comprehensive experiments on multiple public long document summarization datasets from different domains. ATHENA establishes new *state-of-the-art* performance in LRS on all corpora, outperforming previous works significantly.

Our main contributions are as follows:

Abbreviations: LRS, Low-Resource Summarization.

* Corresponding author.
E-mail addresses: gianluca.moro@unibo.it (G. Moro), l.ragazzi@unibo.it (L. Ragazzi).

[1] https://disi-unibo-nlp.github.io/projects/athena/

https://doi.org/10.1016/j.neucom.2023.126356
0925-2312/

- To the best of our knowledge, our paper pioneers the exploration of document representation learning for low-resource summarization by presenting a novel *align-then-abstract learning model* (ATHENA) in which the text segmentation is jointly learned to yield better summaries.
- Our proposed approach *nails multiple issues* related to long-input processing: (i) reduces the GPU memory usage, (ii) avoids input truncation, and (iii) works with few labeled training instances.
- ATHENA achieves new *state-of-the-art* results in low-resource summarization on various well-known datasets.

2. Related work

Fine-tuning pre-trained models for downstream tasks is a standard strategy, but it is often ineffective if solely dozens of labeled training instances are available. Language models are usually pre-trained with self-supervised learning techniques with numerous unlabeled data [2,13–15]. Consequently, downstream-specific pretraining strategies [16] have been introduced to create low-resource-oriented models, e.g., PEGASUS [3]. Nonetheless, with few training samples, pre-trained language models still struggle to adapt to new data from diverse domains [17]. For this reason, several approaches have been proposed to tackle the limited availability of labeled instances. Prompt-based methods [18,19] tune continuous prompts to adapt quickly to new tasks with few examples. Other works [20,21] applied synthetic data augmentation, enhancing the summarization accuracy in low-resource conditions but only experimenting on short texts of max 400 and 200 tokens, respectively. Conversely, to mimic the real-world LRS scenario over long documents, Bajaj et al. [8] proposed an extract-then-abstract approach to provide exclusively salient sentences to a pre-trained model. Despite its effectiveness, this solution involves a two-stage training, in which the summarization model does not process all information in the long input. Finally, Chen and Shuai [22] introduced a meta-transfer learning technique that augments the training data with multiple similar corpora.

Unlike prior contributions, we experiment with multiple public long document summarization datasets from different domains using a base model without synthetic data augmentation.

3. Background

Problem Definition We define the problem of *long document summarization* with the following setup. The input is a lengthy text $\mathcal{X} = \{x_1, \ldots, x_x\}$ coupled with its corresponding summary $\mathcal{Y} = \{y_1, \ldots, y_y\}$, where each $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ is a token. The standard training algorithm adopts the cross-entropy loss, which requires the model to predict the next token y_i of the target summary given $\mathcal{X}$ and the previous target tokens $y_{1:i-1}$, as follows:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{|\mathcal{Y}|} \log p_\theta(y_i | y_{1:i-1}, \mathcal{X}) \tag{1}$$

where θ indicates the model parameters and p is the predicted probability over the vocabulary.

Model Architectures The number of input tokens could be potentially significant (e.g., >10,000). Therefore, $\mathcal{X}$ cannot be processed at its full size with current quadratic transformer-based models and commodity hardware environments. For this reason, several model architectures have been proposed to handle long-input processing: (i) *Efficient sequence-to-sequence* reads more input tokens thanks to sparse attention mechanisms with linear complexity in the input size [11,12,23–25]. (ii) *Extract-then-abstract* lessens the input size by supplying just a subset of the source to the summarization model [8,26,27]. (iii) *Segment-then-abstract* divides the input into sections, independently summarized and concatenated to produce the final summary [9,28–30,42].

4. Method

In this section, we describe our *align-then-abstract representation learning model* (ATHENA) in detail. In a nutshell, ATHENA comprises two key collaborating modules tackling long-input processing (Section 4.1) and document representation learning (Section 4.2). The two components work jointly to learn the best text segmentation that yields better summaries. Our model differs from existing ones thanks to an end-to-end learning solution in which segmentation and summarization cooperate to generate the synthesis. An overview of our proposed and existing architectures is shown in Fig. 1.

4.1. Long-input processing

In long document summarization, the source length, in terms of the number of tokens, may exceed the limit summarization models can consume (e.g., BART [2] truncates inputs longer than 1024 tokens). Nevertheless, we argue that the whole document information can contribute to the final summary. To this aim, our ATHENA model is trained to synthesize a long input by segmenting it into small coherent chunks, learning end-to-end the best document segmentation, thus summarizing the chunks and concatenating the chunk-level summaries to produce the final prediction. Consequently, our model can read the entire document without truncating any information or relying on a subset of snippets.

Source Segmentation To segment a long input into small chunks, we leverage the SE3 algorithm [30]. This unsupervised method uses a BERT-based model to semantically represent the sentences and create the chunks based on their meaning.[2] Unlike SE3, which employs a frozen BERT, we follow the same algorithm but introduce a novel loss (Section 4.2) to train the segmenter end-to-end to uncover the best document segmentation that improves the summarization accuracy. Overall, ATHENA segments a long input $\mathcal{X}$ into n not-overlapping chunks $\{\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_n\}$, each with a number of tokens $\leqslant \mathcal{M}$, corresponding to the max input size summarization models can process (e.g., $\mathcal{M} = 1024$ for BART).

Target Alignment We align each sentence $y_j \in \mathcal{Y}$ to the chunk that can better summarize it, yielding new high-correlated instances $(\bar{\mathcal{X}}, \bar{\mathcal{Y}})$. More precisely, each sentence y_j is assigned to the chunk $\bar{\mathcal{X}}_i$ that maximizes the formula:

$$\mathcal{R}_p^1(\bar{\mathcal{X}}_i, \bar{\mathcal{Y}}_i \circ y_j) \tag{2}$$

where $\mathcal{R}_p^1$ stands for the ROUGE-1 precision metric [31] and $\circ$ indicates the concatenation of previous target sentences already assigned to $\bar{\mathcal{X}}_i$. Note that this matching algorithm does not assure $\bar{\mathcal{Y}}_i \neq \varnothing$, noticeable if the number of chunks exceeds the summary sentences. Yet, we aim to create high-correlated chunk-target pairs $(\bar{\mathcal{X}}, \bar{\mathcal{Y}})$ for the summarization model (see Section 5.8).

4.2. Document representation learning

Our method is trained end-to-end to maximize the conditional probability of generating $\bar{\mathcal{Y}}_i$ from $\bar{\mathcal{X}}_i$, where $(\bar{\mathcal{X}}, \bar{\mathcal{Y}})$ is the set of chunk-target pairs produced by the segmentation and alignment modules. To jointly train the segmenter, we propose *alignment loss* to teach the model to segment better $(\mathcal{X}, \mathcal{Y})$ into more *aligned* pairs $(\bar{\mathcal{X}}, \bar{\mathcal{Y}})$, learning the best document representation.

[2] Please refer to the original paper [30] for details.

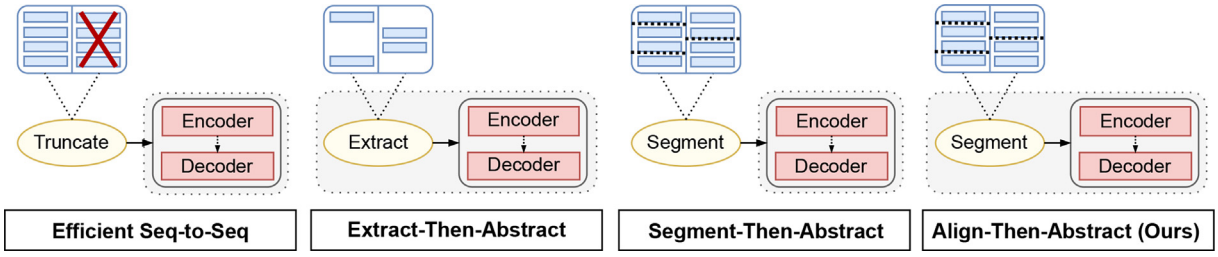


Fig. 1. The establish model architectures for long document summarization. Gray parts indicate the learning modules. The "align-then-abstract" is our proposed architecture.

Alignment Loss The segmenter is trained to maximize the alignment between each chunk-target pair $(\bar{\mathcal{X}}_i, \bar{\mathcal{Y}}_i)$ in terms of semantic content coverage, encouraging the model to locate the best text segmentation that improves the summarization. In detail, the alignment loss $\mathcal{L}^{\eta}_{align}$ computes the cosine similarity between $(\bar{\mathcal{X}}^e, \bar{\mathcal{Y}}^e)$, where e denotes the embedding representation of each $(\bar{\mathcal{X}}_i, \bar{\mathcal{Y}}_i) \in (\bar{\mathcal{X}}, \bar{\mathcal{Y}})$. Specifically, to obtain $\bar{\mathcal{X}}^e$, we generate the sentence embeddings with the segmenter (η denotes its parameters) by computing the mean pooling operation over the token embeddings of each sentence [32]. Afterward, we calculate the mean over the output vectors to obtain a single embedding for each chunk and target. With this loss, the model learns to maximize the alignment between the chunk-target pairs, thus learning to better segment and represent the documents. More precisely, the weights of the model updated during this learning are the same used for segmenting the document sentences into chunks. Hence, we train the model to segment the document better by maximizing the alignment between the chunk-target pairs created after the segmentation. The alignment loss is the following:

$$\mathcal{L}^{\eta}_{align} = 1 - \frac{\bar{\mathcal{X}}^e \cdot \bar{\mathcal{Y}}^e}{|\bar{\mathcal{X}}^e||\bar{\mathcal{Y}}^e|} \tag{3}$$

Summarization Loss The summarization module takes as input the chunk-target pairs and is trained to generate the next output token for each target by minimizing the negative log-likelihood with the following function:

$$\mathcal{L}^{\gamma}_{summ} = -\frac{1}{|\bar{\mathcal{Y}}|}\sum_{t=1}^{|\bar{\mathcal{Y}}|} \log p(y_t | y_{1:t-1}, \bar{\mathcal{X}}) \tag{4}$$

where γ are the parameters of the summarization module, $\bar{\mathcal{X}}$ is the input chunk, and $y_{1:t}$ are the tokens from position 1 to t of the target $\bar{\mathcal{Y}}$. Note that, for the training process, we take only the chunk-target pairs $(\bar{\mathcal{X}}_i, \bar{\mathcal{Y}}_i)$ such that $\bar{\mathcal{Y}}_i \neq \varnothing$. In contrast, we consider all the chunks at inference time.

Algorithm 1: Align-then-abstract learning

Input:
$\mathcal{X} = \{x_1, \ldots, x_x\}$ ▷Input
$\mathcal{Y} = \{y_1, \ldots, y_y\}$ ▷Output
BERT ▷Segmenter
BART ▷Summarizer
Training:
1: $\bar{\mathcal{X}} \leftarrow$ *Segmentation*$(\mathcal{X}, \text{Bert})$ ▷Chunks
2: $\bar{\mathcal{X}}, \bar{\mathcal{Y}} \leftarrow$ *Alignment*$(\bar{\mathcal{X}}, \mathcal{Y})$ ▷Chunk-target pairs
3: *AlignmentLoss*$(\bar{\mathcal{X}}, \bar{\mathcal{Y}}, \text{Bert})$ ▷Backpropagation over BERT
4: *SummarizationLoss*$(\bar{\mathcal{X}}, \bar{\mathcal{Y}}, \text{Bart})$ ▷Backpropagation over BART

4.3. Training objective

The overall training objective of our solution is the following:

$$\mathcal{L}^{\eta,\gamma}(\hat{\mathcal{Y}}_i, \mathcal{Y}_i) = \mathcal{L}^{\eta}_{align} + \mathcal{L}^{\gamma}_{summ} \tag{5}$$

The whole model is trained end-to-end with an align-then-abstract approach to segment the source into content-wise chunks and summarize each of them. Concretely, the segmenter is optimized with the alignment loss, whereas the summarizer is optimized with the summarization loss (Alg. 1). Moreover, we involve an update step of the model weights with a dynamic mini-batch gradient descent equal to the number of chunks per instance, formally defined as follows:

$$\theta_j = \theta_j - \epsilon \frac{1}{n} \sum_{i=n\cdot k}^{n(k+1)} \nabla_{\theta_j} \mathcal{L}(\hat{\mathcal{Y}}_i, \mathcal{Y}_i) \tag{6}$$

where $\theta = \eta + \gamma$ are the parameters of the full model, $n = |\bar{\mathcal{X}}|_i$ is the number of chunks of the i-th document, and $k = \{1, \frac{\mathcal{N}}{n}\}$ is the number of update steps, with $\mathcal{N}$ equal to the amount of training samples. In this way, the gradients are (i) computed for each document chunk, implicitly augmenting the training instances, (ii) averaged per document, and (iii) descended after each instance.

Fig. 2 illustrates our proposed solution. Technically, while the input is a single training instance, the summarizer reads more labeled samples (3 in the example) produced by the alignment module. At inference time, the final summary is obtained by concatenating the predicted chunk-level summaries. Note that the chunks without an assigned target are not considered during training.

5. Experiments

5.1. Setup

To assess the performance of our solution, we simulate a low-resource scenario with labeled data scarcity, adopting the identical experimental setup in previous comparative works on low-resource summarization [3,22] for a fair comparison and reproducibility. Specifically, we train our model with all datasets' first 10 and 100 training instances.

5.2. Datasets

We contemplate the following well-known public long document summarization datasets from different domains and text sizes as evaluation benchmarks. Key measurements are reported in Table 1.[3]

[3] All datasets can be accessed through Hugging Face:https://huggingface.co/datasets

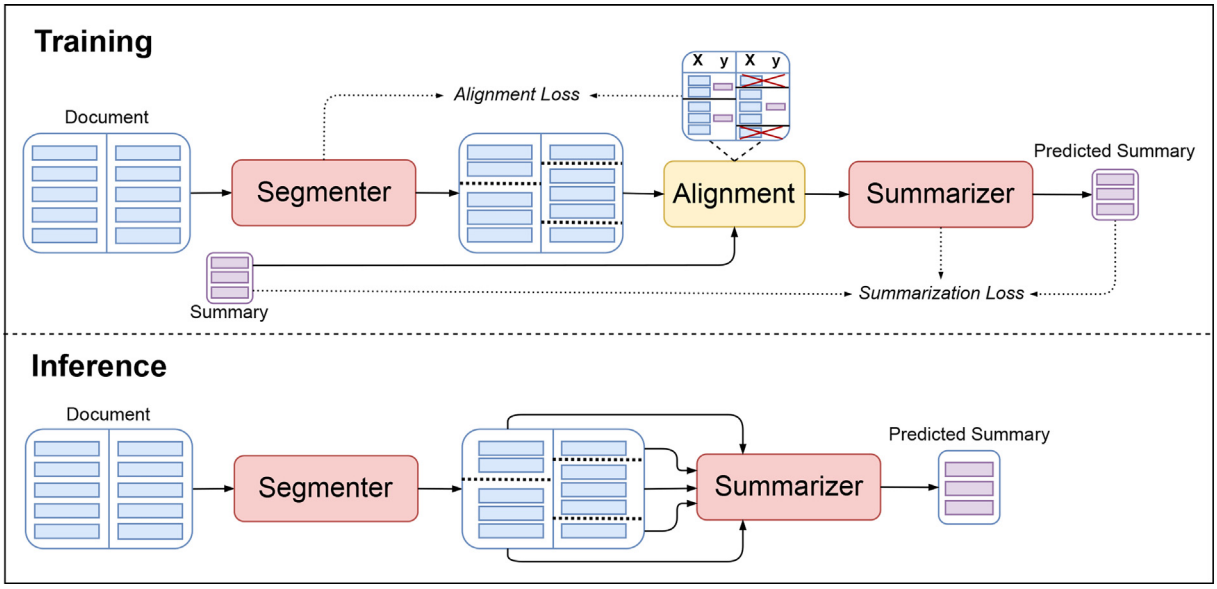


Fig. 2. The illustration of ATHENA at training and inference time. The input is a long document and the output is a short summary.

Table 1
The long document summarization datasets used as testbeds. Statistics include corpus size, number of source texts per instance, number of total words in source and target texts, and source-target coverage, density, and compression ratio of words [33]. Except for the number of samples, all reported values are averaged across all instances.

			Source		Target		Source → Target		
Dataset	**Domain**	**Samples**	Words	Sents	Words	Sents	Coverage	Density	Compress
BILLSUM [34]	Bills	23,455	1673.25	46.49	212.96	4.99	0.90	6.86	12.21
PUBMED [35]	Biomedical	130,397	3166.12	205.39	98.87	7.36	0.89	5.98	16.69
GOVREPORT [24]	Legal	19,463	8765.01	298.69	556.30	18.10	0.94	9.08	17.85

- BILLSUM [34] consists of 22 K U.S. congressional bills from the 103rd-115th (1993–2018) sessions of Congress.
- PUBMED [35] comprises 133 K biomedical publications from PubMed.
- GOVREPORT [24] includes 19 K U.S. government reports.

5.3. Baselines

We compare ATHENA with cutting-edge baselines:

BART [2] is the state-of-the-art denoising sequence-to-sequence pre-trained model for various text generation tasks. Since ATHENA is built upon BART, this comparison can reveal whether the segmentation approach is meaningful in long-input processing. We report the results of BART-base.

PEGASUS [3] is a transformer-based model with a summarization-specific pre-training that helps fast adapting with few labeled data. As there is no public checkpoint of the base version, we include the results of PEGASUS-large.

MTL-ABS [22] is a meta-transfer learning approach for LRS that cope with data scarcity by augmenting the training data with multiple similar corpora.

LED [12] is a state-of-the-art efficient transformer built upon BART with a self-attention mechanism that scales linearly in the input size, allowing the model to process long sequences. We employ LED-base.

LONGT5 [36] is a powerful pre-trained model with sparse attention built upon T5. We use LONGT5-base.

5.4. Evaluation metrics

Although human evaluation is deemed the gold standard for estimating model accuracy, it is prohibitively expensive, and recent research has even shown some shortcomings [37]. For these reasons, we embrace automatic evaluation metrics, assessing the inferred summaries from different perspectives.

Lexical Overlap We use ROUGE-{1,2,L} F1 scores against reference summaries, reporting R-1 and R-2 for informativeness and R-L for fluency.[4] Additionally, inspired by [38], we also compute $\mathcal{R} = \text{avg}(\text{R}-1, \text{R}-2, \text{R}-\text{L})/1 + \sigma_r^2$ to derive an aggregated judgment $\in [0, 1]$ (the higher, the better) that penalizes generations with heterogeneous results across the ROUGE dimensions.

Semantic Similarity We report BERTScore F1 (BS) [39],[5] which computes the contextual similarity between a candidate and its reference summary. We use this metric to assess the model performance in the ablation studies.

5.5. Implementation details

Pre-trained Models The summarizer is initialized with BART-base [2] weights, whereas the segmenter is initialized with a pre-trained sentence embedding model based on BERT-small [13].[6] Technically, the segmenter is based on a siamese network and has

[4] We use ROUGE provided by Hugging Face:https://huggingface.co/metrics/rouge.
[5] We use BERTScore from Hugging Face:https://huggingface.co/metrics/bertscore.
[6] The checkpoint is public in Hugging Face: `sentence-transformers/all-MiniLM-L6-v2`.

Table 2
Low-resource summarization performance of ATHENA on all target datasets. † means that the results are from the original papers. The other models are fine-tuned with the same training details of ATHENA. Best scores on each dataset are bolded.

		BILLSUM				PUBMED				GOVREPORT			
Model	**Sample**	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$
PEGASUS†	10	40.48	18.49	27.27	28.51	33.31	10.58	20.05	21.13	–	–	–	–
	100	44.78	26.40	34.40	35.00	34.05	12.75	21.12	22.47	–	–	–	–
MTL-ABS†	10	41.22	18.61	26.33	28.47	34.08	10.05	18.66	20.73	–	–	–	–
	100	45.29	22.74	29.56	32.24	35.19	11.44	19.89	21.96	–	–	–	–
BART	10	45.59	22.83	29.05	32.19	38.07	12.49	19.91	23.22	48.63	18.48	21.23	28.91
	100	49.74	27.15	32.93	36.27	40.29	13.51	21.85	24.90	43.77	13.39	19.91	25.26
LED	10	45.57	22.89	29.05	32.21	38.43	11.90	19.93	23.13	52.73	18.54	20.73	29.94
	100	48.44	26.62	32.21	35.45	42.07	14.34	22.22	25.86	55.35	20.94	22.02	31.95
LONGT5	10	42.75	18.98	25.31	28.72	36.06	10.54	18.31	21.39	50.12	16.55	19.40	28.04
	100	42.79	19.00	25.31	28.74	36.16	10.65	18.36	21.48	50.06	16.51	19.38	28.00
Athena	10	**47.57**	**24.14**	**30.35**	**33.69**	**40.98**	**13.35**	**21.39**	**24.90**	**54.95**	**19.98**	**22.24**	**31.58**
	100	**51.59**	**29.36**	**35.04**	**38.32**	**42.46**	**14.72**	**22.63**	**26.25**	**56.85**	**22.06**	**23.19**	**33.17**

Table 3
Chunk-target alignment in ATHENA of a random sample from BILLSUM.

Chunk 1

Section 1. short title. this act may be cited as the ""new idea"". sec. 2. clarification that wages paid to unauthorized aliens may not be deducted from gross income. in general. subsection (c) of section 162 of the internal revenue code of 1986 is amended by adding at the end the following new paragraph: wages paid to or on behalf of unauthorized aliens. in general. no deduction shall be allowed under subsection (a) for any wage paid to or on behalf of an unauthorized alien, as defined under section 274a(h)(3) of the immigration and nationality act (8 u. s. c. 1324a(h). *[. . .]*

Target 1

New idea - amends the internal revenue code to deny a tax deduction for wages and benefits paid to or on behalf of an unauthorized alien.

Chunk 2

[. . .]

the commissioner of social security, the secretary of the department of homeland security, and the secretary of the treasury, shall jointly establish a program to share information among such agencies that may or could lead to the identification of unauthorized aliens (as defined under section 274a(h), including any no-match letter, any information in the earnings suspense file, and any information in the investigation and enforcement of section 162(c)(4) of the internal revenue code of 1986. disclosure by secretary of the treasury. in general. subsection (i) of section 6103 of the internal revenue code of 1986 is amended by adding at the end the following new paragraph: payment of wages to unauthorized aliens. upon request from the commissioner of the social security administration or the secretary of the department of homeland security, the secretary shall disclose to officers and employees of such administration or department taxpayer identity information of employers who paid wages with respect to which a deduction was not allowed by reason of section 162(c)(4), and taxpayer identity information of individuals to whom such wages were paid, for purposes of carrying out any enforcement activities of such administration or department with respect to such employers or individuals.". record keeping. paragraph (4) of section 6103(p) of such code is amended by striking "(5), or (7)" in the matter preceding subparagraph (a) and inserting , (7), or (9)", and by striking "(5) or (7)" in subparagraph (f) and inserting "(5), (7), or (9)".

Target 2

directs the commissioner of social security and the secretaries of homeland security and the treasury to jointly establish a program to share information that may lead to the identification of unauthorized aliens. requires the secretary of the treasury to provide taxpayer identity information to the commissioner of social security and the secretary of homeland security on employers who paid nondeductible wages to unauthorized aliens and on the aliens to whom such wages were paid.

Chunk 3

[. . .]

section 401 of the illegal immigration reform and immigrant responsibility act of 1996 is amended by striking the last sentence. application to current employees. voluntary election. the first sentence of section 402(a) of such act is amended to read as follows: "any person or other entity that conducts any hiring in a state or employs any individuals in a state may elect to participate in a pilot program.". benefit of rebuttable presumption. paragraph (1) of section 402(b) of such act is amended by adding at the end the following: "if a person or other entity is participating in a pilot program and obtains confirmation of identity and employment eligibility in compliance with the terms and conditions of the program with respect to individuals employed by the person or entity, the person or entity has established a rebuttable presumption that the person or entity has not violated section 274a(a)(2) with respect to such individuals.".

Target 3

amends the illegal immigration reform and immigrant responsibility act of 1996 to: (1) make permanent the pilot program for verifying the employment eligibility of alien workers, (2) apply such program to current employees in addition to new hires, and (3) establish a rebuttable presumption that employers who participate in the pilot program have not violated the prohibition against continued employment of unauthorized aliens.

Table 4
ROUGE scores of ATHENA using different subsets of the training set.

Dataset	Instances	#1	#2	#3	#4	#5
BILLSUM	10	47.57/24.14/30.35	47.88/24.85/30.81	48.57/25.40/31.07	47.36/23.77/29.56	47.54/24.37/30.29
	100	51.59/29.36/35.04	51.04/28.96/34.87	50.66/28.38/34.42	51.23/28.98/34.69	51.09/28.65/34.76
PUBMED	10	40.98/13.35/21.39	41.00/13.16/20.96	41.53/13.37/21.62	41.80/13.63/21.72	40.61/12.86/21.00
	100	42.46/14.72/22.63	43.09/14.95/22.68	43.41/15.52/23.01	43.00/15.23/22.85	43.39/15.40/22.90
GOVREPORT	10	54.95/19.98/22.24	55.72/20.76/22.74	55.58/20.58/22.46	54.96/20.18/22.11	56.02/20.99/22.63
	100	56.85/22.06/23.19	56.94/22.19/23.23	56.39/21.72/23.05	57.06/22.06/23.19	56.95/22.05/23.25

Table 5
Analyzes on the segmented chunks at inference time. All values are averaged.

Dataset	Instances	# Chunks	Chunk Size
BILLSUM	10	4.9	436.0
	100	4.6	451.0
PUBMED	10	9.4	372.7
	100	9.1	383.2
GOVREPORT	10	24.5	361.1
	100	24.2	365.2

Table 6
Running time and GPU memory requirement of models.

Dataset	Complexity	BART	LED	LONGT5	ATHENA (OURS)
BILLSUM	Time (s)	1.0	2.4	5.0	4.7
	GPU (GB)	5.5	7.5	8.1	5.7
PUBMED	Time (s)	1.2	3.5	6.0	9.0
	GPU (GB)	5.7	11.2	14.1	6.8
GOVREPORT	Time (s)	1.4	7.0	10.2	30.2
	GPU (GB)	6.8	15.9	21.1	7.2

already been fine-tuned using a contrastive learning objective for the semantic textual similarity task.

Training We train on all datasets with mixed precision for 20 epochs, saving the model that perform best on the validation set.[7] We apply gradient checkpointing to save memory, use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and set the learning rate to 3e-5. We consider a chunk size between 256 and 1024 tokens for text segmentation. We eventually set the seed to 42 for reproducibility.

Inference We set the beam width to 5 and use the following summary size (min–max) based on experiments and statistics reported in Table 1: BILLSUM (100–300), PUBMED (100–300), GOVREPORT (500–1000). Finally, we utilize an n-grams penalty of 5 for GOVREPORT and 3 for the other datasets.

Hardware Each experiment is run on a single RTX 3090 GPU of 24 GB memory with PyTorch [40] in a workstation with 64 GB of RAM and an Intel®Core™i9-10900X1080 CPU @ 3.70 GHz.

5.6. Results and discussion

The evaluation results are reported in Table 2. ATHENA achieves new state-of-the-art ROUGE scores in low-resource conditions with a wide margin on BILLSUM, PUBMED, and GOVREPORT.

Capacity to summarize long inputs The results suggest the effectiveness of the align-then-abstract approach for long document summarization in low-resource conditions. Indeed, segmenting a long input into small content-wise chunks allows existing models to summarize very lengthy documents by processing small chunks, extending the input size that the model can handle. In this way, the summarization phase reads all document details, handling the long input at its full length, avoiding prior input truncation or processing only a subset of pre-selected sentences.

Capacity to adapt to data scarcity We show the high capability of our model to synthesize long sequences in a low-resource data scenario. During the training phase, the segmenter creates small high-correlated chunk-target pairs and feeds the summarization module many high-quality training samples that augment the number of training instances and the model's focus during at learning time. Table 3 illustrates a qualitative example of chunk-target pairs correlation in ATHENA, namely the training data and their labels.

Capacity to handle multiple domains Our model attains new state-of-the-art results on a comprehensive collection of datasets, indicating a high potential of adapting to different dictionaries and language styles.

Efficiency and effectiveness ATHENA is built upon small and base models, so it is memory-efficient and more practical to use in small- and medium-sized organizations that cannot afford high-budget GPU memories (i.e., more than 12 GB). Regardless, despite its small number of parameters, ATHENA achieves state-of-the-art results, proving its effectiveness and a conceivable additional gain with large models and more GB of GPU memory available.

5.7. Subset analysis

Unlike prior works, we believe that the selection of samples plays a vital role in the final results in low-resource regimes. For this reason, we conduct further experiments on multiple subsets of the training sets. Technically, we use the first 5 not-overlapping subsets with 10 and 100 instances within each subset to assess if the performance of our proposed model remains stable or highly depends on the input data. Table 4 reports the high similarity of the results despite the different training subsets.

5.8. Segmentation analysis

We first study how the segmentation affects the number and the size of the chunks at inference time. Table 5 shows the mean number of chunks per corpus, which is really high for the GOVREPORT dataset. We notice that the documents are segmented into chunks of about 400 tokens in length for all datasets, despite the distinct text sizes.

5.9. Complexity analysis

Our model ATHENA has a quadratic memory growth w.r.t. the chunk size. Therefore, the space complexity to summarize the entire input text is $\mathcal{O}(L_c^2)$, where L_c is the max chunk size (or the model max input length). Table 6 reports the running time and the GPU memory requirement of models on all datasets trained on 10 instances for 1 epoch. The more training time required for ATHENA is due to its capability to read all the long document chunk by chunk, allowing a low-memory occupation.

[7] We use only the first 10 validation samples to simulate real-world low-resource conditions.

Table 7
Ablation studies to validate the effectiveness of the full method ($\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{gen}$). Best results are bolded.

	10					100				
	R-1$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$	**BS**$_{f1}$	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$	**BS**$_{f1}$
BILLSUM										
Full	**47.57**	**24.14**	**30.35**	**33.69**	**86.26**	**51.59**	**29.36**	**35.04**	**38.32**	**87.65**
w/o $\mathcal{L}_{align}$	46.98	23.56	29.69	33.08	86.07	51.11	28.92	34.90	37.98	87.59
w/o $\mathcal{L}_{gen}$	42.07	18.95	25.14	28.45	83.97	42.03	18.92	25.16	28.43	83.97
PUBMED										
Full	**40.98**	**13.35**	**21.39**	**24.90**	**83.86**	**42.46**	**14.72**	**22.63**	**26.25**	**84.29**
w/o $\mathcal{L}_{align}$	40.63	12.92	21.03	24.53	83.64	42.30	14.67	22.54	26.15	84.21
w/o $\mathcal{L}_{gen}$	39.87	12.34	20.29	23.85	83.35	39.81	12.31	20.28	23.82	83.34
GOVREPORT										
Full	54.95	19.98	**22.24**	31.58	84.81	56.85	**22.06**	**23.19**	**33.17**	**85.13**
w/o $\mathcal{L}_{align}$	**55.78**	**20.36**	22.23	**31.94**	**84.95**	**56.94**	21.93	23.13	33.13	85.12
w/o $\mathcal{L}_{gen}$	54.50	19.93	21.92	31.33	84.65	54.28	19.67	21.84	31.15	84.61

Table 8
Comparison with SE3 on the BILLSUM dataset (the results are taken from the original paper [30]). Best values are bolded.

	BillSum (10)				BillSum (100)			
Model	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$	**R-1**$_{f1}$	**R-2**$_{f1}$	**R-L**$_{f1}$	$\mathcal{R}$
SE3 (*1024*)	44.37	21.17	27.57	30.74	47.85	26.67	33.36	35.68
SE3 (*512*)	46.58	22.03	28.23	31.93	49.88	26.84	33.33	36.34
SE3 (*256*)	46.50	23.24	28.54	32.44	48.17	26.55	31.51	35.11
SE3 (*128*)	41.48	22.73	26.37	30.00	42.42	25.42	28.98	32.10
ATHENA (*dynamic*)	**47.57**	**24.14**	**30.35**	**33.69**	**51.59**	**29.36**	**35.04**	**38.32**

5.10. Ablation studies

We conduct ablation studies to investigate the effectiveness of the modules of our solution. In detail, we report the performance of ATHENA after removing alignment loss and generation loss. The results are summarized in Table 7.

We notice that excluding the generation loss (w/o $\mathcal{L}_{gen}$) leads to the most significant loss in performance. Nonetheless, we still achieve competitive results, demonstrating the excellent capability of our solution architecture. Training the model without considering the alignment loss (w/o $\mathcal{L}_{align}$) decreases the performance, showing the importance of creating high-correlated samples in low-resource conditions.

5.11. Comparison with SE3

The architecture of our solution is related to but differs significantly from SE3 [30]. The similarity only lies in using the segmentation algorithm to split a long input into multiple chunks. However, unlike our dynamic chunk creation through learning, the segmentation module in SE3 is frozen with pre-selected chunk sizes. Table 8 reports the results of SE3 on the BILLSUM dataset to further prove the importance of learning how to better segment a long input.

6. Conclusion

In this paper, we propose ATHENA, a novel approach for long document summarization in low-resource conditions, namely with just dozens of labeled training instances available, which is a real-world scenario. ATHENA is trained end-to-end on an align-then-abstract representation learning to better segment a long input to (i) create small content-wise chunks processable with fewer memory requirements, (ii) read the long text in its full size, and (iii) create high-correlated samples to augment the data with high-quality source-target instances. We demonstrate the effectiveness of our solution by benchmarking three datasets of different domains, significantly outperforming the current state-of-the-art in low-resource summarization on all datasets.

For future works, we suggest investigating the following approaches to better model chunks creation: (i) memory-based operations [44] from unsupervised approaches for entity relationships acquisition [45,46] and classes extraction [47,52] to avant-garde semantic parsing solutions such as event extraction [48]; (ii) retrieval-enhanced techniques [49]. Lastly, as proposed for communication networks [50,51], tracking and propagating knowledge refinements across sentences could be critical when tackling extended sequences.

Ethics statement

Real-world organizations and research teams with a small amount of labeled document-summary pairs can get the most out of using our solution. However, because of biases in current pre-trained language models [41], humans should supervise the summarization process to guarantee the quality of the generated summaries. For this work, we do not deal with sensible or dangerous data. All datasets and models are publicly available.

CRediT authorship contribution statement

Gianluca Moro: Conceptualization, Methodology, Supervision, Writing - review & editing. **Luca Ragazzi:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

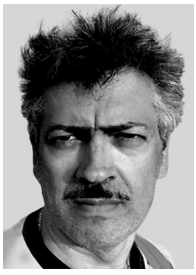
This research is partially supported by (i) the Complementary National Plan PNC-I.1, "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE---DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (ii) the PNRR, M4C2, FAIR---Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program. We thank the Maggioli Group[8] for granting the Ph.D. scholarship to Luca Ragazzi.

References

[1] A.A. Syed, F.L. Gaol, T. Matsuo, A survey of the state-of-the-art models in neural abstractive text summarization, IEEE Access 9 (2021) 13248–13265.
[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 7871–7880. URL:https://aclanthology.org/2020.acl-main.703.10.18653/v1/2020.acl-main.703.
[3] J. Zhang, Y. Zhao, M. Saleh, P.J. Liu, PEGASUS: pre-training with extracted gap-sentences for abstractive summarization, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 11328–11339. URL:http://proceedings.mlr.press/v119/zhang20ae.html.
[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 140 (1–140) (21 2020,) 67.
[5] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, M. Zhou, Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020, volume EMNLP 2020 of Findings of ACL, ACL, 2020, pp. 2401–2410. URL: https://doi.org/10.18653/v1/2020.findings-emnlp.217. 10.18653/v1/2020.findings-emnlp.217.
[6] B. Plank, What to do about non-standard (or non-canonical) language in NLP, in: Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19–21, 2016, volume 16 of Bochumer Linguistische Arbeitsberichte, 2016. URL:https://www.linguistics.rub.de/konvens16/pub/2_konvensproc.pdf.
[7] M.A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, A survey on recent approaches for natural language processing in low-resource scenarios, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Online, 2021, pp. 2545–2568. URL: https://aclanthology.org/2021.naacl-main.201. 10.18653/v1/2021.naacl-main.201.
[8] A. Bajaj, P. Dangati, K. Krishna, P. Ashok Kumar, R. Uppaal, B. Windsor, E. Brenner, D. Dotterrer, R. Das, A. McCallum, Long document summarization in a low resource setting using pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2021, pp. 71–80. URL: https://aclanthology.org/2021.acl-srw.7. 10.18653/v1/2021.acl-srw.7.
[9] J. Wu, L. Ouyang, D.M. Ziegler, N. Stiennon, R. Lowe, J. Leike, P.F. Christiano, Recursively summarizing books with human feedback, CoRR abs/2109.10862 (2021).
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL:https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
[11] M. Zaheer, G. Guruganesh, K.A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, 2020. URL:https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.
[12] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, CoRR abs/2004.05150 (2020).
[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), ACL, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. 10.18653/v1/n19-1423.
[14] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, 2020. URL:https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
[15] K. Clark, M. Luong, Q.V. Le, C.D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. URL:https://openreview.net/forum?id=r1xMH1BtvB.
[16] W. Xiao, I. Beltagy, G. Carenini, A. Cohan, PRIMER: pyramid-based masked sentence pre-training for multi-document summarization, CoRR abs/2110.08499 (2021).
[17] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N.A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 8342–8360. URL:https://aclanthology.org/2020.acl-main.740. 10.18653/v1/2020.acl-main.740.
[18] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021, (Volume 1: Long Papers), Virtual Event, August 1–6, 2021, ACL, 2021, pp. 4582–4597. URL:https://doi.org/10.18653/v1/2021.acl-long.353. 10.18653/v1/2021.acl-long.353.
[19] X. Liu, Y. Gao, Y. Bai, J. Li, Y. Hu, H. Huang, B. Chen, PSP: pre-trained soft prompts for few-shot abstractive summarization, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12–17, 2022, International Committee on Computational Linguistics, 2022, pp. 6355–6368. URL:https://aclanthology.org/2022.coling-1.553.
[20] S. Parida, P. Motlicek, Abstract text summarization: A low resource challenge, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, Hong Kong, China, 2019, pp. 5994–5998. URL:https://aclanthology.org/D19-1616. 10.18653/v1/D19-1616.
[21] A. Magooda, D.J. Litman, Abstractive summarization for low resource data using domain transfer and data synthesis, in: Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference, Originally to be held in North Miami Beach, Florida, USA, May 17–20, 2020, AAAI Press, 2020, pp. 240–245. URL:https://aaai.org/ocs/index.php/FLAIRS/FLAIRS20/paper/view/18441.
[22] Y. Chen, H. Shuai, Meta-transfer learning for low-resource abstractive summarization, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2–9, 2021, AAAI Press, 2021, pp. 12692–12700, URL: https://ojs.aaai.org/index.php/AAAI/article/view/17503.
[23] M. Guo, J. Ainslie, D.C. Uthus, S. Ontañón, J. Ni, Y. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences, CoRR abs/2112.07916 (2021).
[24] L. Huang, S. Cao, N. Parulian, H. Ji, L. Wang, Efficient attentions for long document summarization, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Online, 2021, pp. 1419–1436. URL: https://aclanthology.org/2021.naacl-main.112. 10.18653/v1/2021.naacl-main.112.
[25] J. Phang, Y. Zhao, P.J. Liu, Investigating efficiently extending transformers for long input summarization, CoRR abs/2208.04347 (2022).
[26] Z. Mao, C.H. Wu, A. Ni, Y. Zhang, R. Zhang, T. Yu, B. Deb, C. Zhu, A.H. Awadallah, D.R. Radev, DYLE: dynamic latent extraction for abstractive long-input summarization, CoRR abs/2110.08168 (2021).
[27] G. Moro, L. Ragazzi, L. Valgimigli, D. Freddi, Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Dublin, Ireland, 2022, pp. 180–189. URL:https://aclanthology.org/2022.acl-long.15. 10.18653/v1/2022.acl-long.15.
[28] A. Gidiotis, G. Tsoumakas, A divide-and-conquer approach to the summarization of long documents, IEEE ACM Trans, Audio Speech Lang. Process. 28 (2020) 3029–3040.
[29] Y. Zhang, A. Ni, Z. Mao, C.H. Wu, C. Zhu, B. Deb, A.H. Awadallah, D.R. Radev, R. Zhang, Summn: A multi-stage summarization framework for long input dialogues and documents, CoRR abs/2110.10150 (2021).
[30] G. Moro, L. Ragazzi, Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 11085–11093. URL:https://ojs.aaai.org/index.php/AAAI/article/view/21357.
[31] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL:https://aclanthology.org/W04-1013.
[32] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019, Hong Kong, China, November 3–7, 2019, ACL, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. 10.18653/v1/D19-1410.
[33] M. Grusky, M. Naaman, Y. Artzi, Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, in: Proceedings of the 2018

[8] https://www.maggioli.com/who-we-are/company-profile

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), ACL, New Orleans, Louisiana, 2018, pp. 708–719. URL: https://aclanthology.org/N18-1065. 10.18653/v1/N18-1065.
[34] A. Kornilova, V. Eidelman, BillSum: A corpus for automatic summarization of US legislation, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, ACL, Hong Kong, China, 2019, pp. 48–56. URL:https://aclanthology.org/D19-5406. 10.18653/v1/D19-5406.
[35] A. Cohan, F. Dernoncourt, D.S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), ACL, New Orleans, Louisiana, 2018, pp. 615–621. URL: https://aclanthology.org/N18-2097. 10.18653/v1/N18-2097.
[36] M. Guo, J. Ainslie, D.C. Uthus, S. Ontañón, J. Ni, Y. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences, in: Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10–15, 2022, ACL, 2022, pp. 724–736. URL: https://doi.org/10.18653/v1/2022.findings-naacl.55. 10.18653/v1/2022.findings-naacl.55.
[37] S. Schoch, D. Yang, Y. Ji, "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation, in: Proceedings of the 1st Workshop on Evaluating NLG Evaluation, ACL, Online (Dublin, Ireland), 2020, pp. 10–16. URL:https://aclanthology.org/2020.evalnlgeval-1.2.
[38] G. Moro, L. Ragazzi, L. Valgimigli, Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy, in: Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7–14, 2023, AAAI Press, 2023, pp. 1–9.
[39] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. URL:https://openreview.net/forum?id=SkeHuCVFDr.
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E.Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.
[41] P.P. Liang, C. Wu, L. Morency, R. Salakhutdinov, Towards understanding and mitigating social biases in language models, in: Proc. of the 38th ICML 2021, 18–24 July 2021, Virtual Event, volume 139, PMLR, 2021, pp. 6565–6576. URL: http://proceedings.mlr.press/v139/liang21a.html.
[42] G. Moro, L. Ragazzi, L. Valgimigli, G. Frisoni, C. Sartori, G. Marfia, Efficient memory-enhanced transformer for long-document summarization in low-resource regimes, Sensors 23 (2023) 3542.
[43] G. Moro, N. Piscaglia, L. Ragazzi, P. Italiani, Multi-language transfer learning for low-resource legal case summarization, Artif. Intell. Law 31 (2023).
[44] G. Moro, A. Pagliarani, R. Pasolini, C. Sartori, Cross-domain & in-domain sentiment analysis with memory-based deep neural networks, in: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, September 18–20, 2018, SciTePress, 2018, pp. 125–136. URL: https://doi.org/10.5220/0007239101270138.
[45] G. Frisoni, G. Moro, A. Carbonaro, Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining, in: DATA, SciTePress, 2020, pp. 121–134. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092009636&partnerID=40&md5=27541a3b46d782bb7984eed8ba7fa8a3.
[46] G. Frisoni, G. Moro, Phenomena Explanation from Text: Unsupervised Learning of Interpretable and Statistically Significant Knowledge, in: DATA, volume 1446, Springer, 2020, pp. 293–318.
[47] G. Domeniconi, G. Moro, R. Pasolini, C. Sartori, Cross-domain text classification through iterative refining of target categories representations, in: A. L. N. Fred, J. Filipe (Eds.), KDIR, Rome, Italy, 21 - 24 October, 2014, SciTePress, 2014, pp. 31–42. URL: https://doi.org/10.5220/5550005069400310042https://doi.org/10.5220/0005069400310042.
[48] G. Frisoni, G. Moro, A. Carbonaro, A survey on event extraction for natural language understanding: Riding the biomedical literature wave, IEEE, Access 9 (2021) 160721–160757.
[49] G. Moro, L. Valgimigli, Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature, Sensors21 (2021).
[50] S. Lodi, G. Moro, C. Sartori, Distributed data clustering in multi-dimensional peer-to-peer networks, in: (ADC), Brisbane, 18–22 January, 2010, volume 104 of CRPIT, ACS, 2010, pp. 171–178. URL: http://portal.acm.org/citation.cfm?id=1862264&CFID=17470975&CFTOKEN=71845406.
[51] W. Cerroni, G. Moro, T. Pirini, M. Ramilli, Peer-to-peer data mining classifiers for decentralized detection of network attacks, in: ADC, volume137 of CRPIT, ACS (2013) 101–108.
[52] G. Moro, M. Masseroli, Gene function finding through cross-organism ensemble learning, BioData Min 14 (2021) 14.



Gianluca Moro received the Ph.D. degree in computer science and engineering from the Department of Electronics, Computer Science and Systems of the University of Bologna, Italy, in 1999. He is professor of text mining, data mining and big data analytics at the Department of Computer Science and Engineering of the University of Bologna and head of the research unit in text mining and natural language processing of the Cesena campus. He co-organized several editions of workshops at VLDB and AAMAS, edited five international books and published more than ninety papers, even in top international conferences such as AAAI, IJCAI, EMNLP, ACL, AAMAS, COLING, etc., also winning some best paper awards. He has led national and international projects on data mining and machine learning research topics and collaborates with several research organizations.

Luca Ragazzi received his B.S. and M.S. degrees (with honors) in computer science and engineering. He is a third-year Ph.D. student at the Department of Computer Science and Engineering, University of Bologna, Italy. He has competencies in natural language processing and understanding. He presented several original papers to top-tier international conferences, such as AAAI and ACL, and participated as session chair for AAAI 2023. He is currently working on single and multi-document summarization in low-resource regimes with state-of-the-art language models.