Using Multimodal and Language-Agnostic Sentence Embeddings for Abstractive Summarization

Chaimae Chellaf[1,2], Salima Mdhaffar[1], Yannick Estève[1], Stéphane Huet[1]
[1]LIA - Avignon Université, France
[2]Lundi Matin, France
chaimae.chellaf-el-hammoud@alumni.univ-avignon.fr

Abstract

Abstractive summarization aims to generate concise summaries by creating new sentences, allowing for flexible rephrasing. However, this approach can be vulnerable to inaccuracies, particularly 'hallucinations' where the model introduces non-existent information. In this paper, we leverage the use of multimodal and multilingual sentence embeddings derived from pretrained models such as LaBSE, SONAR, and BGE-M3, and feed them into a modified BART-based French model. A Named Entity Injection mechanism that appends tokenized named entities to the decoder input is introduced, in order to improve the factual consistency of the generated summary. Our novel framework, SBARThez, is applicable to both text and speech inputs and supports cross-lingual summarization; it shows competitive performance relative to token-level baselines, especially for low-resource languages, while generating more concise and abstract summaries.

Keywords: Abstractive summarization, cross-lingual, multimodal, sentence embeddings

1. Introduction

Automatic text summarization plays a critical role in distilling vast amounts of information into concise and informative outputs (Zhang et al., 2025). With the advent of deep learning, neural approaches have become the foundation of modern summarization systems, leveraging large-scale training data and Transformer-based architectures (Vaswani et al., 2017) to model contextual relationships across sequences of text. These models, such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and more recently, large language models (LLMs) like GPT (Radford et al., 2018; Brown et al., 2020) generate summaries by operating at the token level, predicting the next token conditioned on both the input tokens and the previously generated ones. While these models perform well in abstractive summarization, they still exhibit notable limitations that impact the overall quality of the summaries (Shakil et al., 2024).

Extractive summarization focuses on selecting the most important sentences or phrases directly from the source text to form a summary without generating new material. In contrast, abstractive summarization is a more sophisticated technique that produces new sentences to convey the main ideas of the source text; instead of copying, this technique paraphrases and synthesizes the content to present it in a coherent and natural way (Zhang et al., 2025). In addition, two important extensions of text summarization are cross-lingual summarization (Wang et al., 2022) and cross-modal summarization (Sharma et al., 2022; Retkowski et al., 2025). Cross-Lingual Summarization aims to generate summaries in a language different from the source text, while cross-modal summarization involves different input-output modalities, such as speech-to-text or text-to-speech summarization. These approaches expand access to information across languages and modalities. However, both tasks face significant challenges, such as the scarcity of parallel summarization datasets, and error propagation in pipeline systems (Perez-Beltrachini and Lapata, 2021; Wang et al., 2022). For example, speech-to-text summarization is often compromised by transcription errors, while cross-lingual cascade models can be marred by translation errors.

To address these challenges, Barrault et al. (2024) presented Large Concept Models (LCMs) which operate independently of any instantiation in a particular language or modality. LCMs propose a shift from the traditional token-level representation of input texts to sentence-level representation. Instead of processing input text as discrete tokens, they operate within a continuous sentence embedding space wherein each sentence is encoded as a dense vector capturing its underlying semantic content. This representation framework facilitates more abstract reasoning over higher-level semantic units, thereby improving conceptual generalization.

In line with this work, we introduce a novel abstractive summarization model that, unlike LCMs, is designed to work under significantly reduced computational requirements, tailored specifically for the summarization task. While LCM targets general-purpose language modeling, our approach focuses on learning abstract summaries using compact architectures trained on task-specific data. Our con-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9873–9883
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tributions are:

- We propose a novel modified BART-based architecture, SBARThez, which performs summarization from sentence embeddings derived from pretrained models such as LaBSE, SONAR, and BGE-M3. In contrast to conventional token-level encoding methods, our architecture breaks new ground by effortlessly processing text data from various languages as well as audio inputs. We make the training code of our models freely available at `https://github.com/cchellaf/SBARThez`
- We introduce a strategy that incorporates the explicit injection of named entities into the decoder step of the summarization process, aiming to mitigate the risk of hallucinated entities.
- We conduct extensive evaluations of our architecture on a diverse set of benchmarks spanning multiple language pairs (X to French) and two input modalities (text and speech). Our results demonstrate that despite its smaller scale, it achieves competitive performance compared with several baselines, with notable gains on low-resource languages and superior scores on measures of abstractness.

2. Related Work

2.1. Text Sentence Embedding Extraction

Recent advancements in multilingual sentence embeddings have significantly improved the performance of cross-lingual and semantic similarity tasks. One pioneering model, Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), introduced a Siamese network architecture over BERT, enabling efficient computation of semantic similarities between sentence pairs. Initially trained on English, SBERT was later adapted for use in other languages. Another embedding model, LaBSE, introduced by (Feng et al., 2022), uses a novel combination of pretraining and dual-encoder finetuning to enhance translation ranking performance. It offers robust embeddings across 109 languages and has been widely adopted for tasks requiring language-agnostic representations, such as bitext retrieval and downstream classification. More recently, a multilingual and multimodal fixed-size sentence embedding space called SONAR (Duquenne et al., 2023) uses an encoder-decoder approach to build its sentence embedding space, with training on large-scale text data. The SONAR text encoder covers 200 languages, improving both alignment and semantic coherence across these languages. SONAR demonstrated strong performance, particularly in zero-shot cross-lingual retrieval. Jina Embeddings (Sturua et al., 2024) further extended multilingual capabilities by offering embedding models optimized for query-document retrieval, clustering, classification, and text matching, achieving competitive performance across multiple benchmarks. Finally, another notable embedding model, BGE-M3, has emerged as a versatile solution (Chen et al., 2024), supporting over 100 languages and excelling across three key dimensions: multilinguality, multifunctionality, and multigranularity.

2.2. Speech Utterance Embedding Extraction

Compared to earlier approaches, which focused on acoustic frame-level representations (10–20 ms), speech utterance embedding models take a more global approach, learning multimodal and multilingual representations at the speech level (5–10 seconds). These models aim to align the resulting embedding space semantically across languages. For instance, SAMU-XLSR (Khurana et al., 2022) successfully adapted the XLS-R speech encoder (Babu et al., 2022) to generate utterance-level speech embeddings that rival those of language-agnostic sentence embedding models like LaBSE (Feng et al., 2022). This is achieved through a teacher-student framework, where LaBSE is frozen as the teacher, and the speech encoder is trained to generate embeddings that closely match the textual embeddings using cosine similarity. A similar approach is taken in SONAR (Duquenne et al., 2023), where a pretrained sentence embedding space is used as the teacher. This space is constructed using an encoder-decoder model initialized from the NLLB 1B model (Costa-Jussà et al., 2022). The SONAR speech encoders are then trained to align with this shared space, starting from a pretrained w2v-BERT 2.0 model (Barrault et al., 2023). A more recent model, SENSE (Mdhaffar et al., 2025), also leverages w2v-BERT 2.0 as the speech encoder and BGE-M3 (Chen et al., 2024) as the teacher model. These models provide rich, multilingual, and semantically aligned speech representations suitable for downstream tasks like speech summarization.

2.3. Abstractive Summarization

The evolution of neural architectures, particularly sequence-to-sequence (seq2seq) models (Shi et al., 2021) and transformer-based approaches, has revolutionized abstractive text summarization. The availability of large-scale supervised summarization datasets has enabled deep learning techniques to thrive (Grusky et al., 2018; Scialom et al., 2020). Notable milestones include the introduction of self-attention mechanisms in models like BERT (Devlin et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020), which have significantly improved text summarization capabilities. Pretrained large language models, including GPT (Radford et al., 2018), LLaMA (Touvron et al., 2023), and Gemini models (Team et al., 2023), particularly their instruction-tuned variants, have demonstrated better generalization in many natural language processing tasks, including text summarization. They excel at distilling key information into fluent summaries, as shown on recent benchmarks (Zhang et al., 2024).

3. Methodology

We first describe our model's architecture in subsection 3.1 and the pretrained models on which it builds in subsection 3.2. The evaluation metrics used to assess model performance is introduced in subsection 3.3. We then present the summarization tasks and datasets considered in this work in subsection 3.4. Finally, subsection 3.5 outlines the training details of our approach.

3.1. Model Architecture

We propose an abstractive summarization model that generates summaries based on sentence-level embeddings, as illustrated in Figure 1. The input document D is first divided into sentences $D = [s_1, s_2, s_3, \dots, s_n]$. Each sentence s_i is encoded into a vector v_i using a text sentence embedding model. The resulting sequence of vectors $V = [v_1, v_2, v_3, \dots, v_n]$ is then fed into a modified pretrained seq2seq model, which is trained to produce the final summary. To guarantee compatibility with the input requirements of the seq2seq model, we incorporate an additional linear projection layer that dynamically adjusts the dimensionality of sentence embeddings. This layer is only applied when the existing embeddings do not conform to the expected input shape of the model.

In this study, we employ a pretrained token-based encoder-decoder seq2seq model and modify its architecture to accept sentence embedding vectors directly. To achieve this, we remove the encoder's embed-tokens layer, which generates token-level embeddings. By doing so, our approach can leverage sentence-level representations, obviating the need for tokenization within the encoder. As a result, the modified encoder operates on entire sentences, while retaining the unchanged decoder architecture and token-based processing.

The training process of our summarization model consists of two stages. In the first stage, the model is trained on a large-scale textual summarization dataset to adjust its weights to effectively process sentence embeddings. In the second stage, the model is fine-tuned on a dataset aligned with the intended summarization task. Throughout both stages, the text sentence embedding model remains frozen, and only the projection layer and the modified pretrained seq2seq model are trained.

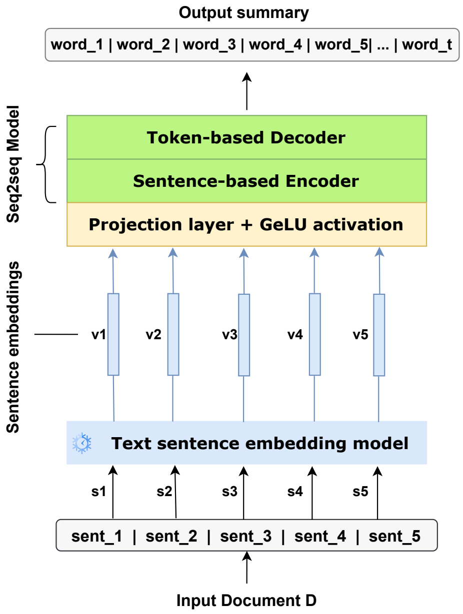

Figure 1: The sentence-based summarization model architecture

3.2. Pretrained Models

Text sentence embedding models: Three embedding models are compared in this study: BGE-M3 (Chen et al., 2024) producing a 1024-dimensional embedding based on the CLS token representation; LaBSE (Feng et al., 2022), which produces a 768-dimensional vector with the same CLS token approach; and SONAR (Duquenne et al., 2023), producing a 1024-dimensional embedding through mean-pooling of its token representations.
Speech utterance embedding models: to validate our method in the speech modality, we integrate it with three distinct speech embedding models. First, we employ the SAMU-XLSR model (Khurana et al., 2022), which leverages LaBSE as its teacher during the training. Next, we utilize SONAR, which supports both text and speech modalities, using its French speech encoder in this context. Lastly, we make use of SENSE (Mdhaffar et al., 2025), which is trained using BGE-M3 as a teacher.
Modified seq2seq model: To achieve our primary objective of creating an abstractive summarization model for French, we chose BARThez (Eddine et al., 2021), a large-scale pretrained seq2seq model specifically tailored for French. With its BART-based architecture and a low computational footprint of 165 M parameters, BARThez has demonstrated strong performance in various

generative tasks. However, we encountered an obstacle when attempting to integrate our sentence embedding models with BARThez: the expected input size for the BART model is 768 dimensions, whereas our embedded representations vary in size. To overcome this issue, we introduced a linear projection layer followed by a GeLU activation function to normalize the inputs. Notably, LaBSE embeddings are exempt from this transformation due to their pre-aligned dimensionality.

In the following, our model will be referred to as **SBARThez** for Semantic-BARThez. To further distinguish between variants that incorporate different sentence embedding models, we will append a suffix to the base model name (e.g., SBARThez-BGE).

3.3. Evaluation Metrics

Since our summarization method relies on high-level semantic abstraction, assessing its quality is challenging: one must gauge the depth of understanding and insight it conveys, not merely its surface-level presentation. To address this, we use various automatic evaluation metrics that range different facets of summarization quality. We report the standard ROUGE-L (Lin, 2004) and BertScore (Zhang et al., 2019) metrics. We also consider three supplementary metrics to assess the abstractiveness of the automatically generated summaries:

ROUGE-1 precision with source document (P-R1) (Lin, 2004) measures the proportion of unigrams (individual words) in the summary that also appear in the source document.

Extractive Fragment Coverage (EFC) (Grusky et al., 2018) evaluates how much of the summary is directly derived from the source text. Extractive fragments are defined as the set of shared sequences of tokens in the source document and the summary. This set is denoted $\mathcal{F}(D,S)$, where D is the document and S is the summary. EFC is calculated by the proportion of summary words that belong to these shared extractive fragments:
EFC$(D,S) = \frac{1}{|S|}\sum_{f\in\mathcal{F}(D,S)}|f|$.

Extractive Fragment Density (EFD) (Grusky et al., 2018) measures the average length of the extractive fragment that each word in the summary belongs to, taking into account both the inclusion of numerous original words and the impact of word arrangement on meaning preservation. Like EFC, EFD is based on extractive fragments, but uses a square of the fragment length:
EFD$(D,S) = \frac{1}{|S|}\sum_{f\in\mathcal{F}(D,S)}|f|^2$.

Named Entity Hallucination Risk (NEHR), introduced by Akani et al. (2023), is calculated as the proportion of entities in the generated summary that are not present in the source document. This metric is used in our study in order to estimate the possible hallucination involving Named Entities.

3.4. Tasks and Datasets

Our proposed summarization approach involves a multi-stage training. The first stage uses the French MLSUM dataset (Scialom et al., 2020), one of the largest news summarization datasets available in French. This step is essential for helping the model adapt to sentence embeddings as input. The second stage focuses on task-specific training, where we tailor our model's performance for three summarization tasks:

Monolingual Text Summarization (Text FR→Text FR): to assess our model's performance on this task, we take the model initially trained on MLSUM and further fine-tune it on the OrangeSum corpus (Eddine et al., 2021).

Cross-lingual Text Summarization (Text X→Text FR): to validate the cross-lingual effectiveness of our approach, the MLSUM-trained model is fine-tuned separately on the French-WikiLingua (Ladhak et al., 2020) and French-CrossSum (Bhattacharjee et al., 2023) datasets, and evaluated on the corresponding cross-lingual test-sets.

Speech Summarization (Speech FR→Text FR): to test our approach on speech, the MLSUM-trained model is further fine-tuned on DECODA (Bechet et al., 2012), which is a French human-human spoken conversation dataset. Each conversation consists of three elements: the audio recording, the manual transcript and a summary. Building upon the work of (Akani et al., 2024), we use the same version of DECODA to ensure comparability.

Table 1 shows the distribution of text-summary pairs across training, development, and test sets of the studied datasets. The statistics for MLSUM, WikiLingua, and CrossSum correspond to their French subsets (French documents paired with their corresponding French summaries).

Dataset	Train	Dev	Test
MLSUM	392,876	16,059	15,828
OrangeSum	21,401	1,500	1,500
DECODA	1,390	396	200
WikiLingua	43,423	6,193	12,405
CrossSum	8,648	1,083	1,083

Table 1: Number of documents per dataset

3.5. Training Details

For each task, we follow the same two-stage learning strategy. In the first training stage, the model is optimized using the AdamW optimizer, with a batch size of 16 and different learning rates for the pretrained seq2seq model and the projection layer. The pretrained seq2seq model parameters were optimized with a learning rate of 1×10^{-5}, while the linear projection layer was trained with a larger learning rate of 1×10^{-3}, allowing faster adaptation of the randomly initialized weights. A weight decay of 1×10^{-5} was applied uniformly to regularize both components. We used the standard cross-entropy loss to train the model on this sequence prediction task. In the second stage, we further fine-tune the model obtained at the end of the first stage.

We present the results of our experiments in the following sections: section 4 covers monolingual text summarization, section 5 addresses cross-lingual text summarization, and section 6 focuses on speech summarization.

4. Monolingual Text Summarization

4.1. Preliminary Results

We first evaluate the originally fine-tuned BARThez model on the OrangeSum test-set. This model serves as our baseline, representing the standard encoder-decoder architecture relying on tokenized input sequences. As shown in Table 2, the SBARThez-BGE model achieves the best performance among all SBARThez variants. Despite its overall improved performance and results that are close to those of the baseline BARThez model in terms of Rouge-L and BertScore, the NEHR remains high for all SBARThez variants. For example, SBARThez-BGE reaches a NEHR of 58.52%, indicating that more than half of its named entities are hallucinated. This limitation is likely due to the highly semantic and abstractive nature of the SBARThez approach, which does not explicitly preserve lexical forms. To address this issue, we suggest including a Named Entity Injection (NEI) mechanism in the SBARThez architecture, as detailed in the following section.

Model	R-L ↑	BertS ↑	NEHR ↓
Ground Truth	100	100	34.01
BARThez	**22.57**	**27.95**	**26.88**
SBARThez-BGE	19.24	24.61	58.52
SBARThez-LaBSE	17.53	22.83	71.79
SBARThez-SONAR	19.10	24.58	59.23

Table 2: Results on the OrangeSum test-set. Best results are **bold**, and second-best results are underlined. BertS refers to BertScore.

4.2. Named Entity Injection Mechanism

To address the issue of hallucinated named entities, we incorporate a NEI mechanism into our approach, as illustrated in Figure 2. Named entities are extracted from the input document using a Named Entity Recognition (NER) system, then tokenized using the seq2seq model's tokenizer and appended to the decoder inputs. During training in both stages, the named entity tokens are incorporated into the decoder block of the BARThez seq2seq model, alongside the standard decoder's input tokens and the encoder's hidden states. At inference time, the decoder is initialized with these tokens as start tokens, allowing the generative process to utilize them in conjunction with the sentence-level embeddings of the input document. This allows the summarization model to leverage both the semantic sentence embeddings on the encoder side and the explicit named entities on the decoder side. The NER model employed in our study is camembert-ner[1], a French NER model fine-tuned from CamemBERT (Martin et al., 2020) on the wikiner-fr dataset (Nothman et al., 2013). The NER system categorizes extracted entities into four classes: PER (persons), ORG (organizations), LOC (locations), and MISC (miscellaneous). Following the extraction, we apply a preprocessing step where entities with a confidence score below 0.9 are filtered out.

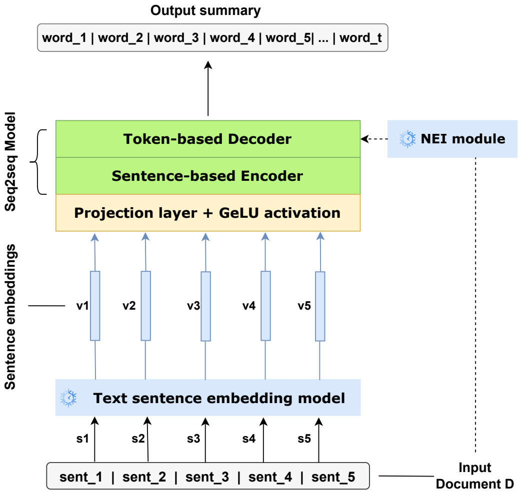

Figure 2: The sentence-based summarization model architecture with the NEI mechanism

4.3. Results using the NEI Mechanism

To assess the effectiveness of the NEI module, we conduct a series of experiments to identify the best named entity (NE) prompt for injection. Among the prompts tested, we identified one structure, which achieved the highest BertScore on OrangeSum: “$(E_1, T_1), (E_2, T_2), \ldots, (E_n, T_n)$.” This prompt lists each named entity E_i alongside its type T_i, where $T_i \in \{\mathrm{PER}, \mathrm{ORG}, \mathrm{LOC}, \mathrm{MISC}\}$, with duplicates removed.It is then tokenized using the BARThez tokenizer to produce NE tokens for the decoder input.

---

[1] `https://huggingface.co/Jean-Baptiste/camembert-ner`

Model	R-L ↑	BertS ↑	NEHR ↓
Ground Truth	100	100	34.01
BARThez	**22.57**	**27.95**	**26.88**
SBARThez-BGE	20.22	26.05	34.16
SBARThez-LaBSE	18.56	24.24	38.41
SBARThez-SONAR	19.40	25.37	36.31

Table 3: Results on the OrangeSum test-set, with SBARThez models using the NEI module.

As shown in Table 3, NE injection improves the performance of all variants of our approach compared to the results presented in Table 2. The best-performing model remains the token-based BARThez model, followed by SBARThez-BGE with NE injection, which achieves the highest scores among all SBARThez variants.

Regarding named entity hallucination, the ground truth summaries exhibit a NEHR of 34.01%, indicating that 34% of the named entities in the summaries do not appear in the original documents. This suggests that human paraphrasing during summary creation may be a contributing factor, such as using "France" to refer to "French Republic". It also highlights a limitation of NEHR, since it relies on exact matching and is therefore overly strict. Nevertheless, it provides a useful proxy for estimating the risk of named entity hallucination.

BARThez achieves the lowest NEHR, indicating stronger faithfulness to the source text in terms of NE generation. However, this approach comes at the cost of limiting the model's ability to generalize or rephrase entities more naturally and expressively. While this may enhance factual accuracy, it may also restrict the model's creative potential. In contrast, our SBARThez models enhanced with NE injection show a significant and consistent reduction in named entity hallucination. Specifically, SBARThez-BGE originally hallucinated over half of the named entities (Tab 2), but with the addition of the NE injection module, the NEHR decreased to 34%, comparable to the ground truth.

4.4. Analysis of Abstractiveness

Table 4 presents the results of the SBARThez variants with NE injection. All token-based encoding models are fine-tuned on OrangeSum. The evaluation includes Rouge-L and BertScore, as well as the copy rate metrics to assess the abstractiveness of the generated summaries.

The SBARThez-BGE model outperforms other variants in terms of Rouge-L and BertScore, while also surpassing mT5-small. Its performance is remarkably close to that of mBART-50 and BARThez, which lead on these metrics. One notable aspect of the token-level encoding models is their extremely high P-R1, EFC, and EFD values, indicating an over-reliance on copying from the source document. For example, the P-R1 value shows that BARThez uses over 90% of the words in its generated summaries that are present in the source document. In contrast, SBARThez models exhibit lower copying rates and demonstrate a remarkable ability to match human summaries in abstractiveness while maintaining competitive performance on R-L and BertScore metrics. Besides, SBARThez summaries tend to be shorter than ground-truth summaries, according to the calculated length. This suggests the model's ability to condense complex information into concise summaries.

Model	Ground Truth		Source Document			Length
	R-L ↑	BertS ↑	P-R1 ↓	EFC ↓	EFD ↓	
Token-Encoding						
BARThez	22.57	**27.95**	91.65	89.11	6.66	24.71
flanT5-base	**22.61**	27.54	94.00	92.37	8.60	27.87
mT5-small	20.22	24.06	95.22	95.39	9.99	25.88
mBART-50	22.05	26.76	93.13	90.96	8.63	27.72
Sentence-Encoding						
SBARThez-BGE	20.22	26.05	78.54	72.30	1.67	20.67
SBARThez-LaBSE	18.56	24.24	**74.64**	**68.71**	**1.45**	21.90
SBARThez-SONAR	19.40	25.37	78.52	72.70	1.60	20.93
Ground Truth	100.0	100.0	78.36	72.84	2.41	32.13

Table 4: Comparison of summarization methods on OrangeSum in terms of abstractiveness, accuracy, and summary length (average words excluding punctuation).

5. Cross-lingual Text Summarization

5.1. Evaluating Model Robustness across Multiple Languages

In this section, we evaluate the cross-lingual abstractive summarization capabilities of our approach. The goal is to generate French summaries from source documents written in various languages. To this end, we fine-tuned the MLSUM-trained model on the French WikiLingua dataset, which contains source documents and summaries in French. This finetuning improves the model's relevance to the dataset's specific characteristics. The NE injection module is not used, as WikiLingua consists of how-to guides, which contain very few named entities. Indeed, our analysis of the training set source documents revealed an average of only 1.37 named entities per document.

For evaluation, we used a unified WikiLingua test-set constructed as the intersection of test sets of the studied source languages (French, English, Spanish, Italian, Portuguese, Dutch, German), aligned through their corresponding French summaries, resulting in 2,166 common test samples.

We observe from table 5 that performance across languages for each SBARThez model remains relatively consistent. For example, the BertScore is around 28 for EN, ES, IT, PT and NL with SBARThez-BGE, indicating that summarization quality does not significantly vary between these languages. The same trend holds for the other two SBARThez variants. However, when comparing these results to the FR→FR setup, we notice a drop in performance. This may be because SBARThez models were fine-tuned on the French WikiLingua dataset, making them more suitable for processing text embeddings derived from French data.

	SBARThez BGE		SBARThez LaBSE		SBARThez SONAR	
	R-L ↑	**BertS ↑**	**R-L ↑**	**BertS ↑**	**R-L ↑**	**BertS ↑**
FR	24.18	32.80	20.17	26.32	23.18	32.21
EN	20.29	28.30	18.87	24.89	21.56	30.03
ES	20.06	28.70	18.80	24.25	21.54	29.62
IT	19.89	28.63	18.25	24.38	20.82	29.47
PT	19.70	28.50	18.56	25.22	21.19	30.07
NL	20.23	28.35	19.16	25.38	21.94	30.10
DE	19.52	27.76	18.71	25.16	21.71	30.47

Table 5: Results on WikiLingua commun test-set.

5.2. Performance in High and Low-Resource Language Settings

As in the previous section, the goal here is to generate French summaries from source documents written in both high- and low-resource languages. We use the MLSUM-trained model, retraining it solely on the French portion of CrossSum (French documents paired with French summaries) to improve its relevance to the dataset's specific requirements. Due to the scarcity of accurate NER systems for low-resource languages, we opt out of using the NE injection module. We employ SBARThez-BGE and compare its performance against two baselines:

Translate-Then-Summarize (TTS) involves translating source documents into French using the M2M-100 1.2B Machine Translation (MT) model (Fan et al., 2021), followed by summary generation using BARThez fine-tuned on French CrossSUM.

A multilingual large language model (LLM), specifically LLaMA-8B, is prompted in a zero-shot setting to produce French summaries.

The results are presented in Tables 6 and 7. The evaluation set, denoted as Test-Dev, combines the test and development sets. The number of samples is displayed alongside each result. Our approach excels in high-resource languages according to BertScore, outperforming other methods in English, Ukrainian and Japanese while closely matching the TTS method's performance in other high-resource languages. Notably, our SBARThez-based method outperforms both TTS and LLM approaches across all low-resource languages, with exceptions for Urdu (UR) and Sinhala (SI).

A key point to emphasize is that languages such as Igbo (IG), Kirundi (RN), and Pidgin (PCM) were not seen in the training of the BGE embedding model. Despite this limitation, our SBARThez-BGE approach still achieves strong results and outperforms all other methods on these languages. In contrast, for Kirundi and Pidgin, we were unable to find suitable MT systems to translate these languages into French, rendering the TTS approach ineffective in the absence of available MT models.

Overall, these findings demonstrate the significant advantage of our approach for cross-lingual summarization, particularly in low-resource languages, where it maintains strong performance even in the absence of NER resources or embedding model support for the source language.

6. Speech Summarization

6.1. Experimental Setup and Results

Three experiments were conducted to adapt the MLSUM-trained model to the speech modality. First, the model was fine-tuned on manual transcriptions from DECODA, resulting in a text-based model. Second, we fine-tuned the model on DECODA speech data, using three separated speech utterance embeddings generated by either SONAR, SAMU-XLSR, or SENSE. Finally, we experimented with fine-tuning the model using both speech and text modalities by merging and shuffling their text and speech embeddings.

Baselines: For comparative analysis, we also fine-tuned BARThez on the manual transcriptions of DECODA and integrated it with two different Automatic Speech Recognition (ASR) systems. **Cascaded_1** uses Whisper Tiny for ASR and BARThez for summarization, resulting in a Word Error Rate (WER) of 64.13% on the test set. In contrast, **Cascaded_2** leverages Whisper Large for ASR and BARThez for summarization, achieving a WER of 26.01% on the test set.

NE extraction: To enable the NE injection module, we fine-tuned the Whisper-small model (Radford et al., 2023) on DECODA to minimize WER. The resulting ASR model was integrated with the SpeechBrain toolkit (Ravanelli et al., 2024) to transcribe speech and then extract named entities from the transcriptions using the NER model.

Audio segmentation: To segment the audio files and extract the sequence of speech utterance embeddings used as input to the SBARThez model, we rely on the ground-truth time boundaries provided in DECODA. Each audio file in the corpus is divided into segments with annotated start and end times. Accordingly, if an audio file contains 8 segments, we generate 8 corresponding speech embeddings; one for each segment.

Results Analysis: Results presented in Table 8 reveal that all SBARThez variants exhibit their best performance on the speech test-set when trained with both text and speech modalities. This indicates

	EN	AR	RU	UK	KO	JA	VI	ID	SW	HI
SBARThez-BGE (ours)	**29.40**	29.83	29.15	**29.91**	28.59	**33.25**	28.92	30.64	30.95	25.61
MT+BARThez	29.11	**30.71**	**30.05**	28.89	**28.68**	15.65	**30.18**	**31.62**	**31.33**	**26.49**
LLaMA-8B	25.00	26.33	24.61	25.61	24.19	29.23	26.12	25.61	21.48	20.68
#Test-Dev Samples	377	197	160	143	52	54	88	204	174	119

Table 6: Results on CrossSum using BertScore ↑ – **High-Resource Languages**

	PA	UR	PS	MR	YO	SO	AZ	TA	MY	SI	GU	IG	RN	PCM
SBARThez-BGE (ours)	**18.40**	17.19	**28.38**	**25.81**	**30.75**	**28.52**	**29.44**	**25.72**	**26.03**	25.98	**25.11**	**26.77**	**26.14**	**29.81**
MT+BARThez	16.05	17.67	27.42	24.37	16.25	05.28	22.54	15.24	06.44	**27.09**	03.09	05.45	-	-
LLaMA-8B	14.62	**21.61**	25.99	19.89	23.96	21.64	24.22	15.32	21.65	24.03	19.38	17.92	22.57	25.53
#Test-Dev Samples	45	152	36	46	33	100	42	84	17	19	46	20	144	88

Table 7: Results on CrossSum using BertScore ↑ – **Low-Resource Languages**.

that training with speech embeddings enhances the model's performance on speech. In addition, SBARThez-BGE achieves a BertScore of 35.07 on the text test-set and 35.05 on the speech test-set, demonstrating consistent performance across modalities. As shown in the same table, WER has a clear impact on the performance of cascaded approaches. For example, when switching from Whisper Large (lower WER) to Whisper Tiny, the BertScore drops significantly, from 38.54 to 32.55. In comparison, while SBARThez-BGE and SBARThez-SONAR outperform the Cascaded_1, they are still behind the Cascaded_2 system. This performance gap may be explained by the use of a generic NER system that can only extract PER, ORG, and LOC entities, while DECODA mainly contains domain-specific entities such as phone numbers, prices, and transport types, which are not captured by the generic NER system.

Model	**Train Data**	**Test Data**	**R-L ↑**	**BertS ↑**
BARThez	Text	Text	29.11	38.90
Cascaded_1	-	Speech	23.43	32.55
Cascaded_2	-	Speech	**27.82**	**38.54**
SBARThez BGE	Text	Text	24.64	35.07
	Text	Speech	22.86	32.05
	Speech	Speech	23.99	34.64
	Both	Speech	24.43	35.05
SBARThez LaBSE	Text	Text	22.26	32.39
	Text	Speech	21.03	30.28
	Speech	Speech	21.09	31.03
	Both	Speech	22.05	31.39
SBARThez SONAR	Text	Text	24.52	34.89
	Text	Speech	20.88	32.44
	Speech	Speech	23.58	32.73
	Both	Speech	23.34	33.50

Table 8: Results on the DECODA test-set.

6.2. Impact of Audio Segmentation

In real-world scenarios, the ground-truth (GT) audio segmentations used to obtain the results in subsection 6.1 are not available. Therefore, we use automatic audio segmentation methods and analyze their impact on the effectiveness of speech summarization. These methods include:

Content-aware Segmentation: The boundaries of an audio segment are determined by the acoustic or linguistic properties of the signal. We evaluate two methods: **(1) Voice Activity Detection (VAD)** detects the presence or absence of human speech, segmenting the audio where speech begins and ends. We use the Silero-VAD tool (Team, 2024). **(2) Speaker Diarization (Spk DIAR)** divides audio into homogeneous segments based on speaker identity, thereby segmenting the audio at speaker turns. We use the pyannote toolkit (Bredin, 2023).

Time-driven Segmentation: The audio is divided into fixed-length segments, regardless of speech or speaker content. This simple approach relies solely on duration. We evaluate three variants with segment lengths of 3, 5, and 8 seconds.

We evaluate the fine-tuned SBARThez-BGE model, trained on manual transcriptions of DECODA, using different segmentation strategies (Table 9). The results show that all segmentation methods maintain summary quality, with Rouge-L and BertScore values comparable to those obtained using GT segmentation. Among automatic methods, segmentation into 5-second segments achieves the best performance after the GT segmentation, while 8-second segments also yield results close to the GT baseline. This indicates that our approach remains effective even when using longer audio segments. Overall, our method demonstrates consistent performance across various segmentation approaches, including simple time-based ones, making it valuable for real-world applications.

Category	**Method**	**R-L ↑**	**BertS ↑**
GT Segmentation	-	*24.64*	*35.07*
Content-aware Segmentation	VAD	23.51	34.43
	Spk DIAR	24.18	34.95
Time-driven Segmentation	3sec fixed	23.83	34.23
	5sec fixed	**24.42**	**34.98**
	8sec fixed	23.80	34.22

Table 9: Segmentation results with SBARThez-BGE. Evaluation is made on speech segments using different segmentation methods.

7. Conclusion

In this paper, we introduce SBARThez, a novel abstractive summarization approach that leverages semantic sentence-level embeddings, moving beyond traditional token-level encoding. We address the issue of hallucinated named entities with a Named Entity Injection mechanism. Our experimental results show that SBARThez produces more abstractive summaries than token-based models, while maintaining competitive performance and surpassing strong baselines in the cross-lingual setting, particularly for low-resource languages. Our approach is also highly versatile: it can be applied to both text and speech inputs, and supports multilingual source documents. This allows for efficient generation of high-quality summaries from multiple languages without requiring additional training or adaptation. This work opens up promising avenues for future research in end-to-end speech summarization and cross-lingual summarization.

8. Limitations

Our study focuses on generating French abstractive summaries. While the methodology could extend to other languages, including English or low-resource ones, further validation is needed to confirm its generalizability. Furthermore, although we demonstrate the effectiveness of our approach for summarization, its applicability to other NLP tasks, such as question answering, paraphrasing, or text simplification, remains to be explored. Generalizing the method across different tasks could provide broader insights into its versatility and robustness. Finally, one key limitation in this study is the lack of suitable factual consistency metrics for French. Existing open-source metrics are primarily designed for English, and while G-Eval (Liu et al., 2023) is currently one of the best-performing metrics with GPT-based models, its reliance on closed-source, high-cost models prevented us from using it in our evaluation. Despite these limitations, we believe this work introduces a promising new approach to abstractive summarization and makes a meaningful contribution to ongoing research in the field.

9. Acknowledgments

This work used HPC resources from GENCI-IDRIS: grants AD011015509R1, A0181012551, and AD011016301.

10. Bibliographical References

Eunice Akani, Frédéric Bechet, Benoît Favre, and Romain Gemignani. 2024. Unified framework for spoken language understanding and summarization in task-based human dialog processing. In *INTERSPEECH 2024*.

Eunice Akani, Benoit Favre, Frederic Bechet, and Romain Gemignani. 2023. Reducing named entity hallucination risk to ensure faithful summary generation. In *16th International Natural Language Generation Conference*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. XLS: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv e-prints*, pages arXiv–2412.

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. 2012. DECODA: a call-centre human-human spoken conversation corpus. In *LREC*, volume 1, pages 1343–1347.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564.

Hervé Bredin. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, volume 1 (long and short papers)*, pages 4171–4186.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained french sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen Mckeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Salima Mdhaffar, Haroun Elleuch, Chaimae Chellaf, Ha Nguyen, and Yannick Estève. 2025. Sense models: an open source solution for multilingual and multimodal semantic-based tasks. In *ASRU 2025*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Jan Niehues, and Alexander Waibel. 2025. From speech to summary: A comprehensive survey of speech summarization. *arXiv preprint arXiv:2504.08024*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, page 128255.

Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. 2022. End-to-end speech summarization using restricted self-attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076. IEEE.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2(1):1–37.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Comput. Surv.*, 57(11).

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.