Supervising the Centroid Baseline for Extractive Multi-Document Summarization

Simão Gonçalves[♋] **Gonçalo Correia**[♋] **Diogo Pernes**[♋♭] **Afonso Mendes**[♋]
[♋]Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[♭]Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
{simao.goncalves, goncalo.correia, diogo.pernes, amm}@priberam.pt

Abstract

The centroid method is a simple approach for extractive multi-document summarization and many improvements to its pipeline have been proposed. We further refine it by adding a beam search process to the sentence selection and also a centroid estimation attention model that leads to improved results. We demonstrate this in several multi-document summarization datasets, including in a multilingual scenario.

1 Introduction

Multi-document summarization (MDS) addresses the need to condense content from multiple source documents into concise and coherent summaries while preserving the essential context and meaning. Abstractive techniques, which involve generating novel text to summarize source documents, have gained traction in recent years (Liu and Lapata, 2019; Jin et al., 2020; Xiao et al., 2022), following the advent of large pre-trained generative transformers. However, their effectiveness in summarizing multiple documents remains challenged. This is attributed not only to the long input context imposed by multiple documents but also to a notable susceptibility to factual inconsistencies. In abstractive methods, this is more pronounced when compared to their extractive counterparts due to the hallucination-proneness of large language models.

Extractive approaches, on the other hand, tackle this problem by identifying and selecting the most important sentences or passages from the given documents to construct a coherent summary. Extractive MDS usually involves a sentence importance estimation step (Hong and Nenkova, 2014; Cao et al., 2015; Cho et al., 2019), in which sentences from the source document are scored according to their relevance and redundancy with respect to the remaining sentences. Then, the summary is built by selecting a set of sentences achieving high relevance and low redundancy. The centroid-based method (Radev et al., 2000) is a cheap unsupervised solution in which each cluster of documents is represented by a centroid that consists of the sum of the TF-IDF representations of all the sentences within the cluster and the sentences are ranked by their cosine similarity to the centroid vector. While the original method is a baseline that can be easily surpassed, subsequent enhancements have been introduced to make it a more competitive yet simple approach (Rossiello et al., 2017; Gholipour Ghalandari, 2017; Lamsiyah et al., 2021).

In this work, we refine the centroid method even further: i) we utilize multilingual sentence embeddings to enable summarization of clusters of documents in various languages; ii) we employ beam search for sentence selection, leading to a more exhaustive exploration of the candidate space and ultimately enhancing summary quality; iii) we leverage recently proposed large datasets for multi-document summarization by adding supervision to the centroid estimation process. To achieve this, we train an attention-based model to approximate the oracle centroid obtained from the ground-truth target summary, leading to significant ROUGE-score improvements in mono and multilingual settings. To the best of our knowledge, we are the first to tackle the problem within a truly multilingual framework, enabling the summarization of a cluster of documents in different languages.[1]

2 Related Work

Typical supervised methods for extractive summarization involve training a model to predict sentence saliency, i.e. a model learns to score sentences in a document with respect to the target summary, either by direct match in case an extractive target is available or constructed (Svore et al., 2007; Woodsend and Lapata, 2012; Mendes et al., 2019) or by maximizing a similarity score (*e.g.*, ROUGE)

[1] https://github.com/Priberam/cera-summ

Proceedings of The 4th New Frontiers in Summarization Workshop, pages 87–96

with respect to the abstractive target summaries (Narayan et al., 2018). Attempts to reduce redundancy exploit the notion of maximum marginal relevance (MMR; Carbonell and Goldstein, 1998; McDonald, 2007) or are coverage-based (Gillick et al., 2008; Almeida and Martins, 2013), seeking a set of sentences that cover as many concepts as possible while respecting a predefined budget. During inference, the model is then able to classify the sentences with respect to their salience, selecting the highest-scored sentences for the predicted summary. Rather than training a model that predicts salience for each individual sentence, we employ a supervised model that directly predicts an overarching summary representation, specifically predicting the centroid vector of the desired summary. Training this model can thus be more direct when training with abstractive summaries (as is the case in most summarization datasets), since computing the reference summary centroid is independent of whether the target is extractive or abstractive.

Regarding enhancements to the centroid method for extractive MDS, Rossiello et al. (2017) refined it by substituting the TF-IDF representations with word2vec embeddings (Mikolov et al., 2013), and further incorporated a redundancy filter into the algorithm. Gholipour Ghalandari (2017), on the other hand, retained the utilization of TF-IDF sentence representations but improved the sentence selection process. Recently, Lamsiyah et al. (2021) introduced modifications to the sentence scoring mechanism, incorporating novelty and position scores, and evaluated a diverse array of sentence embeddings with the proposed methodology, including contextual embeddings provided by ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019).

While there have been initiatives to foster research in multilingual extractive MDS (Giannakopoulos, 2013; Giannakopoulos et al., 2015), the proposed approaches (Litvak and Vanetik, 2013; Aries et al., 2015; Huang et al., 2016) are only language-agnostic, requiring all the documents within each cluster to be in the same language. In contrast, we address extractive MDS in a scenario where each cluster is multilingual.

3 Methodology

The pipeline of our proposed model is divided into two stages. In the first stage, we use an attention model to obtain a cluster representation that replaces the naive centroid obtained by averaging sentence embeddings of the documents in a cluster. The rationale behind this approach is that the contribution of each sentence to the cluster centroid should depend on its relevance to the cluster summary. In order to capture the whole cluster context, a sentence-level attention model is employed, assigning variable weights to each sentence embedding so as to approximate the resulting average to the centroid that would be obtained by averaging the sentence embeddings of the target summary. In the second stage, an adapted version of the greedy sentence selection algorithm from Gholipour Ghalandari (2017) for extractive MDS is used to select the sentences included in the predicted summary. This adapted version uses our proposed supervised centroid and also includes a beam search algorithm to better explore the space of candidate summaries.

3.1 Centroid Estimation

Gholipour Ghalandari (2017) builds a centroid by summing TF-IDF sentence representations of all the sentences that compose the cluster to summarize. In our research, we compute the centroid from a learnable weighted average of the contextual sentence embeddings, via an attention model.

Attention Model In our centroid estimation procedure, we use a pre-trained multilingual sentence transformer from Yang et al. (2020) to encode the sentences from the news articles, obtaining contextual embeddings $\boldsymbol{e}_k \in \mathbb{R}^d$, $k \in \{1, \ldots, N\}$, for each of the N sentences in a cluster. Since it is often the case that the first sentences of a document are especially important for news summarization tasks, we add sentence-level learnable positional embeddings to the contextual embeddings at the input of the attention model. Specifically, given a cluster D comprising N sentences, we compute:

$$\boldsymbol{e}_{\text{pos},k} = \boldsymbol{e}_k + \boldsymbol{p}_{\text{pos}(k)}, \tag{1}$$

where $\text{pos}(k)$ is the position within the respective document of the k-th sentence in the cluster and $\boldsymbol{p}_{\text{pos}(k)} \in \mathbb{R}^d$ is the corresponding learnable positional embedding. Each $\boldsymbol{e}_{\text{pos},k} \in \mathbb{R}^d$ is then concatenated with the mean-pool vector of the cluster,[2] denoted by $\overline{\boldsymbol{e}_{\text{pos}}} \in \mathbb{R}^d$, resulting in $\boldsymbol{e}'_{\text{pos},k} = \text{concat}(\boldsymbol{e}_{\text{pos},k}, \overline{\boldsymbol{e}_{\text{pos}}})$ for each sentence. This concatenation ensures that the computation of

[2]This is calculated by averaging the sentence embeddings within each document and then computing the mean of these individual document averages.

the attention weight for each position uses information from all the remaining positions. The vector $\boldsymbol{\beta} \in \mathbb{R}^N$ of attention weights is obtained as:

$$\boldsymbol{\beta} = \text{softmax}\left(\text{MLP}(\boldsymbol{e}'_{\text{pos},1}), \ldots, \text{MLP}(\boldsymbol{e}'_{\text{pos},N})\right), \tag{2}$$

where MLP is a two-layer perceptron shared by all the positions. It has a single output neuron and a hidden layer with d units and a tanh activation.

After computing the attention weights for the cluster, we take the original sentence embeddings $\boldsymbol{e}_k$, $k \in \{1, \ldots, N\}$, and compute a weighted sum of these representations:

$$\boldsymbol{h} = \sum_{k=1}^{N} \beta_k \boldsymbol{e}_k. \tag{3}$$

Consequently, the resultant vector $\boldsymbol{h} \in \mathbb{R}^d$ is a convex combination of the input sentence embeddings. Since it is not guaranteed that the target centroid lies within this space, $\boldsymbol{h}$ is subsequently mapped to the output space through a linear layer, yielding an estimate $\hat{\boldsymbol{c}}_{\text{attn}} \in \mathbb{R}^d$ of the centroid. Hereafter we refer to this attention model as **Ce**ntroid **R**egression **A**ttention (CeRA).

Interpolation The original (unsupervised) approach involves estimating the centroid by computing the average of all sentence representations $\boldsymbol{e}_k$ within a cluster, which has consistently demonstrated strong performance. Let $\overline{\boldsymbol{e}}_D$ represent this centroid for cluster D. To leverage the advantages of this effective technique, we introduce $\overline{\boldsymbol{e}}_D$ as a residual component to enhance the estimate produced by the attention model. Thus, our final centroid estimate is computed as:

$$\hat{\boldsymbol{c}} = \boldsymbol{\alpha} \odot \hat{\boldsymbol{c}}_{\text{attn}} + (1 - \boldsymbol{\alpha}) \odot \overline{\boldsymbol{e}}_D, \tag{4}$$

where $\boldsymbol{\alpha} \in [0, 1]^d$ is a vector of interpolation weights and $\odot$ denotes elementwise multiplication. The interpolation weights are obtained from concatenating $\hat{\boldsymbol{c}}_{\text{attn}}$ and $\overline{\boldsymbol{e}}_D$ and mapping it through an MLP of two linear linear layers with d units each. The two layers are interleaved with a ReLU activation and a sigmoid is applied at the output. We call the model with interpolation CeRAI.

Training Objective Finally, we minimize the cosine distance between the model predictions $\hat{\boldsymbol{c}}$ and the mean-pool of the sentence embeddings of the target summary $\boldsymbol{c}_{\text{gold}}$.

3.2 Sentence Selection

Considering the cluster D and a set S with the current sentences in the summary. at each iteration of greedy sentence selection (Gholipour Ghalandari, 2017), we have

$$\boldsymbol{e}_{S \cup \{s\}} = \sum_{s' \in S} \boldsymbol{e}_{s'} + \boldsymbol{e}_s \tag{5}$$

for each sentence $s \in D \setminus S$. Then, the new sentence s^* to be included in the summary is

$$s^* = \arg\max_{s \in D \setminus S} \cos\text{sim}(\boldsymbol{e}_{S \cup \{s\}}, \overline{\boldsymbol{e}}_D), \tag{6}$$

where $\cos\text{sim}$ is the cosine similarity. The algorithm stops when the summary length reaches the specified budget.[3] As demonstrated in that work, redundancy is mitigated since the centroid is compared to the whole candidate summary $S \cup \{s\}$ at each iteration and not only to the new sentence s.

In our version of the algorithm, we not only estimate the cluster centroids as explained in §3.1, replacing $\overline{\boldsymbol{e}}_D$ by $\hat{\boldsymbol{c}}$ in equation (6), but also employ a beam search (BS) algorithm so that the space of candidate summaries is explored more thoroughly. Moreover, in order to exhaust the chosen budget, we add a final greedy search to do further improvements to the extracted summary. The procedure is defined in Algorithm 1, shown in Appendix A, and we describe it less formally below.

Beam Search The process begins by preselecting sentences, retaining only the first n sentences from each document. Beam search initiates by selecting the top B sentences with the highest similarity scores with the centroid, where B represents the beam size. In each subsequent iteration, the algorithm finds the highest-scoring B sentences on each beam, generating a total of B^2 candidates. Among these candidates, only the highest-ranked B sentences are retained. Suppose any of these sentences exceed the specified budget length for the summary. In that case, we preserve the corresponding previous state, and no further exploration is conducted on that beam. The beam search concludes when all candidate beams have exceeded the budget or when no more sentences are available.

Greedy Search To exhaust the specified budget and improve results, we add a greedy search of

[3] While the original algorithm would stop after the first sentence that exceeded the budget, we stop before it is exceeded, and thus we do not need truncation to respect the budget.

Method	Multi-News	WCEP-10	TAC2008	DUC2004
Oracle centroid	21.72 ± 0.33	28.54 ± 1.21	11.99 ± 1.32	10.29 ± 1.01
Gholipour Ghalandari	16.07 ± 0.26	15.09 ± 0.92	7.36 ± 1.15	6.82 ± 0.76
Lamsiyah et al.	13.92 ± 0.22	16.10 ± 0.96	7.91 ± 1.31	**7.80** ± 0.78
BS (*Ours*)	16.22 ± 0.25	15.64 ± 0.97	8.10 ± 1.32	7.03 ± 0.64
BS+GS (*Ours*)	16.70 ± 0.26	16.41 ± 0.91	8.16 ± 1.25	7.46 ± 0.83
CeRA (*Ours*)	17.98 ± 0.23	**17.46** ± 0.98	8.27 ± 1.26	7.31 ± 0.74
CeRAI (*Ours*)	**17.99** ± 0.27	17.24 ± 0.93	**8.37** ± 1.24	7.72 ± 0.77

Table 1: ROUGE-2 recall with 95% bootstrap confidence intervals of different extractive methods on the considered test sets. CeRA and CeRAI were only trained on the Multi-News training dataset.

sentences that are allowed within the word limit. The top-scoring B states from the beam search are used as starting points for this greedy search. Then, for each state, we greedily select the highest-scoring sentence that does not exceed the budget among the top T ranked sentences. This process iterates until either all of the top T ranked sentences would exceed the budget or there are no further sentences left for consideration.

4 Experimental Setup

Herein, we outline the methods, datasets, and evaluation metrics employed in our experiments.

Methods We compare our approaches with the centroid-based methods from Gholipour Ghalandari (2017) and Lamsiyah et al. (2021), described in §2. To be consistent with the remaining methods, the approach by Gholipour Ghalandari (2017) was implemented on top of contextual sentence embeddings instead of TF-IDF. Additionally, we perform ablation evaluations in three scenarios: i) a scenario (BS) where we do not use the centroid estimation model (§3.1) and rely solely on the beam search for the sentence selection step (§3.2); ii) a scenario (BS+GS) identical to the previous one, except that we perform the greedy search step after the beam search; iii) two scenarios (CeRAI and CeRA) where we utilize the centroid estimation model with and without incorporating interpolation, and apply the BS+GS algorithm on the predicted centroid. The "Oracle centroid" upperbounds our approaches, since it results from applying BS+GS on the mean-pool of the sentence embeddings of the target summary, $\boldsymbol{c}_{\text{gold}}$, as the cluster centroid. Appendix C provides additional details about data processing and hyperparameters.

Datasets We used four English datasets, Multi-News (Fabbri et al., 2019), WCEP-10 (Ghalandari et al., 2020; Xiao et al., 2022), TAC2008, and DUC2004, and one multilingual dataset, CrossSum (Bhattacharjee et al., 2023), in our experiments. We used the centroid-estimation models trained on Multi-News to evaluate CeRA and CeRAI on WCEP-10, TAC2008, and DUC2004 since these datasets do not provide training splits. CrossSum was conceived for single-document cross-lingual summarization, so we had to adapt it for multilingual MDS. This adaptation results in clusters that encompass documents in multiple languages, with each cluster being associated with a single reference summary containing sentences in various languages. We explain this procedure and provide further details about each dataset in Appendix B.

Evaluation Metrics We evaluate ROUGE scores (Lin, 2004) in all the experiments. When evaluating models in the multilingual setting, we translated both the reference summaries and the extracted summaries into English prior to ROUGE computation. As we optimized for R2-R on the validation sets, we report it as our main metric in Tables 1 and 2. The remaining scores are shown in Appendix D.

5 Results

Monolingual Setting The ROUGE-2 recall (R2-R) of all the methods in the monolingual datasets are presented in Table 1. F1 scores and results for the other ROUGE variants are presented in Table 4, in Appendix D. The first observation is that BS alone outperforms Gholipour Ghalandari (2017) in all datasets, with additional improvements obtained when the greedy search step is also performed (BS+GD). This was expected since our approach explores the candidate space more thoroughly. The motivation for using a supervised centroid estimation model arose from the excellent ROUGE results obtained when using the target summaries to build the centroid ("Oracle centroid" in the tables), showing that an enhanced centroid estimation procedure could improve the results substantially. This is con-

Method	CrossSum	CrossSum-ZS
Oracle centroid	11.74 ± 0.55	14.91 ± 0.49
Gholipour Ghalandari	7.72 ± 0.43	10.03 ± 0.40
Lamsiyah et al.	8.01 ± 0.52	10.45 ± 0.46
BS (*Ours*)	7.74 ± 0.44	10.16 ± 0.40
BS+GS (*Ours*)	8.23 ± 0.43	10.85 ± 0.41
CeRA (*Ours*)	**9.65** ± 0.49	11.67 ± 0.41
CeRAI (*Ours*)	9.38 ± 0.50	**11.73** ± 0.43

Table 2: ROUGE-2 recall results with 95% bootstrap confidence intervals of different extractive methods on the multilingual test sets. The CrossSum set contains the same languages used for training the centroid estimation model, whereas CrossSum-ZS (*zero-shot*) consists of languages that were not present in the training data.

firmed by the two methods using the centroid estimation model (CeRA and CeRAI), which improve R2-R significantly in Multi-News and WCEP-10 and perform at least on par with Lamsiyah et al. (2021) in TAC2008 and DUC2004. It's also worth noting that CeRA and CeRAI were only trained on the Multi-News training set and nevertheless performed better or on par with the remaining baselines on the test sets of the remaining corpora. Incorporating the interpolation step (CeRAI) appears to yield supplementary enhancements compared to the non-interpolated version (CeRA) across various settings, which we attribute to this method adding regularization to the estimation process, improving results on harder scenarios.

Multilingual Setting The R2-R scores of all the methods in CrossSum can be found in Table 2, while additional results are in Table 5 of Appendix D. Once again, we observe the superiority of the centroid estimation models, CeRA and CeRAI, in comparison to all the remaining methods, with the variants with and without interpolation performing on par with each other. Most notably, these models prove to be useful even when tested with languages unseen during the training phase, underscoring their robustness and applicability in a zero-shot setting.

6 Conclusions

We enhanced the centroid method for multi-document summarization by extending a previous approach with a beam search followed by a greedy search. Additionally, we introduced a novel attention-based regression model for better centroid prediction. These improvements outperform existing methods across various datasets, including a multilingual setting, offering a robust solution for this challenging scenario. Regarding future work, we believe an interesting research direction would be to further explore using the supervised centroids obtained by the CeRA and CeRAI models, by having them as a proxy objective to obtain improved abstractive summaries.

Limitations

While we believe that our approach possesses merits, it is equally important to recognize its inherent limitations. Diverging from conventional centroid methods that operate entirely in an unsupervised manner, our centroid estimation model necessitates training with reference summaries. Nevertheless, its robustness to dataset shifts was demonstrated: the model trained on Multi-News consistently yielded strong results when assessed on different English datasets, and the model trained on a subset of languages from CrossSum displayed successful generalization to other languages.

Finally, our method introduces increased computational complexity. This arises from both the forward pass through the attention model and the proposed beam search algorithm, which incurs a greater computational cost compared to the original, simpler greedy approach proposed by Gholipour Ghalandari (2017).

Acknowledgements

This work is supported by the EU H2020 SELMA project (grant agreement No. 957017).

References

Miguel Almeida and André Martins. 2013. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 196–206, Sofia, Bulgaria. Association for Computational Linguistics.

Abdelkrime Aries, Djamel Eddine Zegour, and Khaled Walid Hidouci. 2015. AllSummarizer system at MultiLing 2015: Multilingual single and multi-document summarization. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–244, Prague, Czech Republic. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *preprint arXiv:1607.06450*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin 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, Toronto, Canada. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019. Multi-document summarization with determinantal point processes and contextualized representations. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 98–103, Hong Kong, China. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

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. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *preprint arXiv:2005.10070*.

Demian Gholipour Ghalandari. 2017. Revisiting the centroid-based method: A strong baseline for multi-document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria. Association for Computational Linguistics.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Daniel Gillick, Benoit Favre, and Dilek Hakkani-Tür. 2008. The ICSI summarization system at TAC 2008. In *Proceedings of Text Understanding Conference*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Taiwen Huang, Lei Li, and Yazhao Zhang. 2016. Multilingual multi-document summarization with enhanced hlda features. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 299–312, Cham. Springer International Publishing.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. Association for Computational Linguistics.

Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, and Saïd El Alaoui Ouatik. 2021. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications*, 167:114152.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marina Litvak and Natalia Vanetik. 2013. Multilingual multi-document summarization with POLY2. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 45–49, Sofia, Bulgaria. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.

Afonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. Jointly extracting and compressing documents with summary state representations. 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 3955–3966, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, Valencia, Spain. Association for Computational Linguistics.

Krysta Svore, Lucy Vanderwende, and Christopher Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 448–457, Prague, Czech Republic. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243, Jeju Island, Korea. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

A Sentence Selection Algorithm

Algorithm 1 Sentence Selection

Require: Cluster D, centroid $\hat{\boldsymbol{c}}$, summary budget ℓ, number of sentences n to pre-select, beam size B, number of candidates T for greedy search.

$D_n \leftarrow$ select-first(D, n)
$\pi, \pi_{\text{next}}, \pi_{\text{bs}} \leftarrow$ empty list
while $\exists b : \text{length}(\pi_{\text{next}}[b]) < \ell$ **do**: ▷ Beam Search
  $\pi_{\text{next}} \leftarrow \text{BS}_{\text{step}}(\pi, D_n, B, \hat{\boldsymbol{c}})$ [4]
  if $\exists b : \text{length}(\pi_{\text{next}}[b]) > \ell$ **then**
    π_{bs}.append(π)
  end if
  $\pi \leftarrow \forall \pi_{\text{next}}[b] : \text{length}(\pi_{\text{next}}[b]) \leq \ell$
end while
$\pi_{\text{best}} \leftarrow$ highest-scored B states in π_{bs} (sorted)
for $b = 1, 2, \ldots, B$ **do**: ▷ Greedy Search
  $t \leftarrow 0$
  $D'_n \leftarrow D_n \setminus \pi_{\text{best}}[b]$
  while $t < T$ **do**:
    $s^* \leftarrow \arg\max_{s \in D'_n} \cos\text{sim}(\boldsymbol{e}_{\pi_{\text{best}}[b] \cup \{s\}}, \hat{\boldsymbol{c}})$
    $\pi'_{\text{best}}[b] \leftarrow \pi_{\text{best}}[b] \cup \{s^*\}$
    if $\text{length}(\pi'_{\text{best}}[b]) \leq \ell$ **then**:
      $\pi_{\text{best}}[b] \leftarrow \pi'_{\text{best}}[b]$
      $t \leftarrow 0$
    else:
      $t \leftarrow t + 1$
    end if
    $D'_n \leftarrow D'_n \setminus \{s^*\}$
  end while
end for
return $S \leftarrow$ highest-scored state in π_{best}

B Datasets

We now describe each of the datasets used for evaluation and explain how we have adapted CrossSum for the task of MDS.

Multi-News The Multi-News dataset (Fabbri et al., 2019) is a large-scale dataset for MDS of news articles. It contains up to 10 documents per cluster and more than 50 thousand clusters divided into training, validation, and test splits. There is a single human-written reference summary for each cluster.

WCEP-10 This dataset (Ghalandari et al., 2020; Xiao et al., 2022) consists of short human-written target summaries extracted from the Wikipedia Current Events Portal (WCEP). Each news cluster associated with a certain event is paired with a single reference summary, and there are at most 10 documents per cluster. The dataset comprises 1022 clusters, all of which are used for testing.

TAC2008 This is a multi-reference dataset introduced by the Text Analysis Conference (TAC)[5]. It provides no training nor validation sets and the test set consists of 48 news clusters, each with 10 related documents and 4 human-written summaries as references.

DUC2004 Another multi-reference news summarization dataset[6] designed and used for testing only. It contains 50 clusters with 10 documents and 4 human-written reference summaries each.

CrossSum To assess the performance of the models in a multilingual context, we have adapted the CrossSum dataset (Bhattacharjee et al., 2023) for the task of MDS. Initially designed for cross-lingual summarization, this dataset offers document-summary pairs for more than 1500 language directions. The dataset is derived from pairs of articles sourced from the multilingual summarization dataset XL-Sum (Hasan et al., 2021). Notably, these pairings were established using an automatic similarity metric, resulting in many pairs covering similar topics rather than the exact same stories, rendering it well-suited MDS.

To tailor this dataset for our specific task, we began by selecting the data from a predefined subset of the languages. Subsequently, we aggregated the documents into clusters, taking into account their pairings. For instance, if document A was paired with document B and document B was paired with document C, then A, B, and C would belong to the same cluster. Clusters containing only one document were discarded. For obtaining multilingual reference summaries for each cluster, we interleaved the sentences from the individual summaries until we reached a predefined limit of 100 words. We have built training, validation, and test sets using data in English, Spanish, and French, and another test set using data in Portuguese, Russian, and Turkish to evaluate our model in a zero-shot

[4] BS_{step} denotes a step of the usual beam search algorithm. Details omitted for brevity.

[5] https://tac.nist.gov
[6] https://duc.nist.gov

setting. Statistics about each split are presented in Table 3.

C Experimental Details

Data Processing To ensure a fair comparison, all the models we evaluated used the same sentence representations, specifically, sentence embeddings obtained from the `distiluse-base-multilingual-cased-v2`[7] sentence encoder (Yang et al., 2020).

For monolingual datasets, the documents were split into sentences using `sent_tokenize` from the NLTK library (Bird et al., 2009). For CrossSum, we used `SentSplitter` from the multilingual ICU-tokenizer.[8] Regular expressions were applied to replace redundant white spaces and excessive paragraphs and empty sentences were excluded. Before sentence selection (Algorithm 1), the data goes through a second processing step, during which duplicate sentences and sentences that individually exceed the summary budget are eliminated.

When evaluating models in CrossSum, we translated both the reference summaries and the extracted summaries into English prior to ROUGE computation. All the translations were performed using the M2M-100 12-billion-parameter model (Fan et al., 2021).

The following word-limit budgets were used by all models: 230 words for the Multi-News dataset, 100 words for TAC2008, DUC2004 and CrossSum, and 50 words for WCEP-10.[9]

Hyperparameters The hyperparameters for the beam search-based methods were tuned by running a grid search on the BS+GS approach on the Multi-News validation set. For the number of sentences n, odd numbers from 1 to 9 were tested. For the beam width B values 1,5, and 9 were examined, and regarding the number of candidates T, values 1,5, and 9 were considered. The values that maximized R2-R on this validation set were n = 9, B=5, and T=9. In all of our experiments, these were the values we considered for the parameters. Note that for the BS method only n and B are relevant.

The hyperparameters of the centroid estimation model used in CeRA were obtained by random search on Multi-News. The hyperparameters yielding the highest R2-R score on the validation set for the produced summaries were kept. The CeRAI model was trained using the optimal hyperparameters found for CeRA. The optimal parameters were: *batch size* = 2, *learning rate* = 5×10^{-4}, and *number of positional encodings* = 35. We utilized the Adam optimizer with a multi-step learning rate scheduler configured with *step size* = 3 and $\gamma = 0.1$.

Implementation Details Our CeRA and CeRAI models used early stopping, where the stopping criteria metric was based on R2-R. Layer normalization (Ba et al., 2016) was applied on the input data before adding the positional information to it and before passing the data through the last linear layer that transforms $\boldsymbol{h}$ (equation (3)) into $\hat{\boldsymbol{c}}_{\text{attn}}$ in the CeRA and CeRAI models. We have also normalized the input data to have a unit L2 norm.

D Additional Results

The ROUGE-1/2/L recall and F1 scores obtained by all the methods in the monolingual datasets are shown in Table 4. Table 5 presents the same quantities for the multilingual case.

[7] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[8] https://pypi.org/project/icu-tokenizer

[9] We used ROUGE 1.5.5 toolkit with the following arguments: `-n 4 -m -2 4 -l budget -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a`

Split	Languages	#Clusters	#Docs per cluster	Avg #sentences per doc	Avg #words per summary
Train	en, es, fr	6541	2−10	38.5 ± 28.8	52.5 ± 16.1
Val	en, es, fr	889	2−6	34.4 ± 27.4	52.3 ± 15.5
Test	en, es, fr	853	2−6	36.6 ± 35.4	52.2 ± 16.2
Test-ZS	pt, ru, tr	933	2−5	23.4 ± 21.1	60.2 ± 20.8

Table 3: CrossSum: statistics of each split. Averages are indicated with standard deviations.

Test set	Method	R1-R	R1-F	R2-R	R2-F	RL-R	RL-F
Multi-News	Oracle centroid	54.26	50.36	21.72	20.02	24.33	22.42
	Gholipour Ghalandari	47.91	45.64	16.07	15.16	21.41	20.24
	Lamsiyah et al.	44.91	43.02	13.93	13.18	20.56	19.53
	BS (*Ours*)	48.34	45.81	16.22	15.24	21.34	20.08
	BS+GS (*Ours*)	49.54	45.98	16.70	15.36	21.81	20.08
	CeRA (*Ours*)	50.75	47.07	17.98	16.52	**22.69**	20.86
	CeRAI (*Ours*)	**50.76**	**47.08**	**17.99**	**16.53**	**22.69**	**20.87**
WCEP-10	Oracle centroid	58.72	44.94	28.54	21.50	42.38	31.94
	Gholipour Ghalandari	41.26	35.09	15.09	12.61	29.42	24.86
	Lamsiyah et al.	41.65	**35.62**	16.10	**13.38**	30.53	**25.75**
	BS (*Ours*)	43.48	35.07	15.64	12.42	30.49	24.44
	BS+GS (*Ours*)	46.23	34.72	16.41	12.05	31.85	23.60
	CeRA (*Ours*)	**47.14**	35.23	**17.46**	12.65	**33.03**	24.28
	CeRAI (*Ours*)	46.85	35.17	17.24	12.59	32.81	24.24
TAC2008	Oracle centroid	41.07	42.02	11.99	12.26	20.66	21.11
	Gholipour Ghalandari	32.00	34.38	7.36	7.91	16.64	17.87
	Lamsiyah et al.	31.00	33.75	7.91	**8.65**	16.65	18.16
	BS (*Ours*)	33.93	35.62	8.10	8.53	17.62	**18.50**
	BS+GS (*Ours*)	**35.12**	**35.98**	8.16	8.34	**17.99**	18.40
	CeRA (*Ours*)	34.43	35.07	8.27	8.42	17.35	17.66
	CeRAI (*Ours*)	34.44	35.11	**8.37**	8.52	17.73	18.06
DUC2004	Oracle centroid	39.93	41.10	10.29	10.60	19.48	20.05
	Gholipour Ghalandari	32.82	35.86	6.82	7.48	16.00	17.51
	Lamsiyah et al.	32.81	36.03	**7.80**	**8.61**	16.66	**18.34**
	BS (*Ours*)	34.01	36.20	7.03	7.51	16.35	17.41
	BS+GS (*Ours*)	35.11	36.37	7.46	7.74	**16.98**	17.60
	CeRA (*Ours*)	34.88	36.06	7.31	7.56	16.67	17.23
	CeRAI (*Ours*)	**35.16**	**36.38**	7.72	7.99	16.89	17.48

Table 4: ROUGE-1/2/L recall and F1 results of different extractive methods on the considered monolingual test sets.

Test set	Method	R1-R	R1-F	R2-R	R2-F	RL-R	RL-F
CrossSum	Oracle centroid	46.86	31.85	11.74	7.93	27.64	18.57
	Gholipour Ghalandari	38.64	27.88	7.72	5.56	23.30	16.65
	Lamsiyah et al.	37.89	27.53	8.01	5.77	23.81	17.13
	BS (*Ours*)	39.24	27.83	7.74	5.48	23.60	16.53
	BS+GS (*Ours*)	40.78	27.71	8.23	5.57	24.42	16.39
	CeRA (*Ours*)	**42.45**	**28.89**	**9.65**	**6.52**	**25.64**	**17.27**
	CeRAI (*Ours*)	42.31	28.73	9.38	6.31	25.55	17.15
CrossSum-ZS	Oracle centroid	50.55	37.30	14.91	11.00	28.90	21.08
	Gholipour Ghalandari	41.70	32.65	10.03	7.82	24.52	19.02
	Lamsiyah et al.	41.14	32.39	10.45	8.17	24.81	19.31
	BS (*Ours*)	42.53	32.65	10.16	7.81	24.87	18.90
	BS+GS (*Ours*)	44.36	32.65	10.85	7.99	25.74	18.74
	CeRA (*Ours*)	**45.44**	**33.43**	11.67	8.57	**26.52**	**19.30**
	CeRAI (*Ours*)	45.37	33.38	**11.73**	**8.62**	26.51	19.26

Table 5: ROUGE-1/2/L recall and F1 results of different extractive methods on the considered multilingual test sets.