HISum: Hyperbolic Interaction Model for Extractive Multi-Document Summarization

Mingyang Song
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University
Beijing, China
mingyang.song@bjtu.edu.cn

Yi Feng
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University
Beijing, China
21112027@bjtu.edu.cn

Liping Jing[*]
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University
Beijing, China
lpjing@bjtu.edu.cn

ABSTRACT

Extractive summarization helps provide a short description or a digest of news or other web texts. It enhances the reading experience of users, especially when they are reading on small displays (e.g., mobile phones). Matching-based methods are recently proposed for the extractive summarization task, which extracts a summary from a global view via a document-summary matching framework. However, these methods only calculate similarities between candidate summaries and the entire document embeddings, insufficiently capturing interactions between different contextual information in the document to accurately estimate the importance of candidates. In this paper, we propose a new hyperbolic interaction model for extractive multi-document summarization (HISum). Specifically, HISum first learns document and candidate summary representations in the same hyperbolic space to capture latent hierarchical structures and then estimates the importance scores of candidates by jointly modeling interactions between each candidate and the document from global and local views. Finally, the importance scores are used to rank and extract the best candidate as the extracted summary. Experimental results on several benchmarks show that HISum outperforms the state-of-the-art extractive baselines[1].

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

Extractive Summarization, Multi-document Summarization, Representation Learning, Hyperbolic Deep Learning

ACM Reference Format:
Mingyang Song, Yi Feng, and Liping Jing. 2023. HISum: Hyperbolic Interaction Model for Extractive Multi-Document Summarization. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583197

[*] Corresponding author.
[1] https://github.com/MySong7NLPer/HISum

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583197

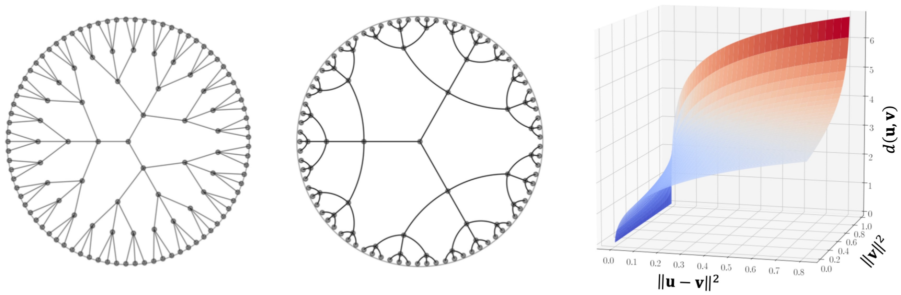
1 INTRODUCTION

Extractive summarization is a fundamental task of Natural Language Processing (NLP) that aims to compress a long document or multi-documents into a shorter version while retaining the most salient information. This is useful in generating a snippet or digest for a searched web page or other web text. Existing summarization methods [16, 33] can be roughly classified into two categories: abstractive and extractive. Concretely, abstractive methods generate a summary word by word, like human writing based on the understanding of the document, making the generated summary more fluent and human-like. Extractive methods extract several sentences from documents as a summary. While abstractive methods can be more concise and flexible, extractive methods can guarantee correct grammar and are more consistent factually [25, 36].

Most existing extractive summarization models typically select saliency from the document at the sentence level by autoregression or non-autoregression strategies. Generally, these methods choose sentences from the source document by modeling relationships between words and sentences [19, 43], sentences and sentences [8, 29, 47], as well as sentences and documents [37, 43] to distinguish the importance score of each sentence to create a summary. However, the above methods of modeling the relationship between sentences are essentially sentence level extractors rather than considering the semantics of the full summary. This makes them more inclined to select highly generalized sentences while ignoring the coupling of multiple sentences [46].

Considering the above issue, MatchSum [46] conceptualizes extractive summarization as a semantic matching problem. MatchSum first extracts multiple salient sentences from the document and creates a set of candidate summaries by combining them. It then globally selects the best candidate by using a document-summary matching framework. Concretely, a Siamese-BERT architecture is used to generate semantically meaningful text embeddings in Euclidean space. These embeddings are compared using cosine similarity to estimate the importance of candidates. The fundamental idea behind this is that a good summary should be more semantically similar to the document than a bad summary.

Although the above models have proved successful for the extractive summarization task, it still suffers from an inherent limitation: the ability to consider the latent hierarchical structures when obtaining text representation and calculating the semantic similarities is bounded by the nature of flat geometry provided by the euclidean space [10, 31, 40]. As mentioned, natural languages often exhibit inherent hierarchical structures ingrained with complex syntax and semantics. Therefore, the latent hierarchical structures should



(A) Distance in Euclidean Space (Euclidean Tree)

(B) Distance in Poincaré Ball (Poincaré Tree)

(C) Growth of Poincaré Distance

Figure 1: Visualization of different spaces. Comparison between trees embedded in Euclidean and Hyperbolic spaces. We adopt geodesics, the analogy of straight lines in Hyperbolic spaces, to link nodes in (B). Line/geodesic segments connecting nodes are approximately the same length in their corresponding spaces. Intuitively, nodes embedded in Euclidean space look more "crowded", while the Hyperbolic space allows sufficient capacity to embed trees and enough distances between leaf nodes. Specifically, (C) denotes the growth of the Poincaré distance $d(\mathrm{u}, \mathrm{v})$ relative to the Euclidean distance and the norm of v (for fixed $||\mathbf{u}|| = 0.9$).

be considered when representing text information and calculating text semantic similarities. Fortunately, the hyperbolic space is naturally suitable for modeling this kind of latent hierarchical structure (e.g., tree-likeness) [48] and is applied in many downstream natural language processing tasks [5, 6, 31]. Therefore, Song et al. [39] proposes a hyperbolic document-summary matching model, which learns text representation and semantic similarity in the hyperbolic space rather than the Euclidean space. However, these methods only consider estimating the importance of each candidate summary from a global view while ignoring the importance of local context, resulting in the extraction of unqualified summaries.

In this paper, we propose a novel hyperbolic interaction model for extractive multi-document summarization (HISum), where local and global interactions are jointly captured in the same hyperbolic space. From a global view, we calculate the text semantic similarities between candidate summaries and the whole document via the Poincaré distance as traditional matching-based models do. In terms of the local view, we first adopt a set of hyperbolic CNN encoders on the document to derive n-gram phrase representations and obtain phrase-aware summary and document representations via the Poincaré distance. And then, we adopt a cosine-similarity layer to get the fine-grained semantic similarity as the local-view interaction. Finally, we combine global and local interactions as the importance scores of candidate summaries and learn to rank the scores via two margin-based triplet loss functions to extract the best candidate. We conduct extensive experiments on several benchmark datasets, prevailing and widely used for the extractive summarization task. Extensive experimental results demonstrate that our model HISum outperforms recent state-of-the-art extractive summarization baselines.

Our contributions can be summarized as follows:

- We propose a new hyperbolic interaction model for extractive multi-document summarization (HISum) and further explore the potential of hyperbolic deep learning in the extractive summarization task.
- HISum first obtains document and candidate summary representations in the same hyperbolic space. Then, jointly modeling interactions from global and local views via the Poincaré distance in the same hyperbolic space to adequately obtain information from the context of the document.
- Extensive experimental results on several benchmark datasets demonstrate that our model HISum outperforms the recent state-of-the-art extractive summarization models.

2 BACKGROUND

Hyperbolic geometry is one type of non-Euclidean geometry with a constant negative curvature, which is regarded as a special case in the Riemannian geometry [18]. Before introducing our model, this section briefly gives the basic information about hyperbolic space. In a traditional sense, hyperbolic spaces are not vector spaces; one cannot use standard operations such as summation, multiplication, etc. To remedy this problem, one can utilize the formalism of Möbius gyrovector spaces allowing one to generalize many standard operations to hyperbolic spaces [21]. Similarly to the previous studies [14, 31, 40], we adopt the Poincaré ball and use an additional hyper-parameter c which modifies the curvature of Poincaré ball; it is then defined as $\mathbb{M}_c^n = \{\mathbf{x} \in \mathbb{R}^n : c||\mathbf{x}||^2 < 1, c \geq 0\}$. The corresponding conformal factor now takes the form $\lambda_{\mathbf{x}}^c := \frac{2}{1-c||\mathbf{x}||^2}$ and c is the curvature of the hyperbolic space. We restate the definitions of fundamental mathematical operations for the generalized Poincaré ball model and refer readers to Ganea et al. [14] for more details. Next, we present details of the closed-form formulas of several Möbius operations.

Möbius Addition. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{M}_c^n$, their addition is defined as,

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c||\mathbf{y}||^2)\mathbf{x} + (1 - c||\mathbf{x}||^2)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c^2||\mathbf{x}||^2||\mathbf{y}||^2}. \tag{1}$$

Möbius Matrix-vector Multiplication. For a linear transformation $\mathbf{M} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\forall \mathbf{x} \in \mathbb{M}_c^n$, if $\mathbf{M}\mathbf{x} \neq 0$, then the *Möbius*

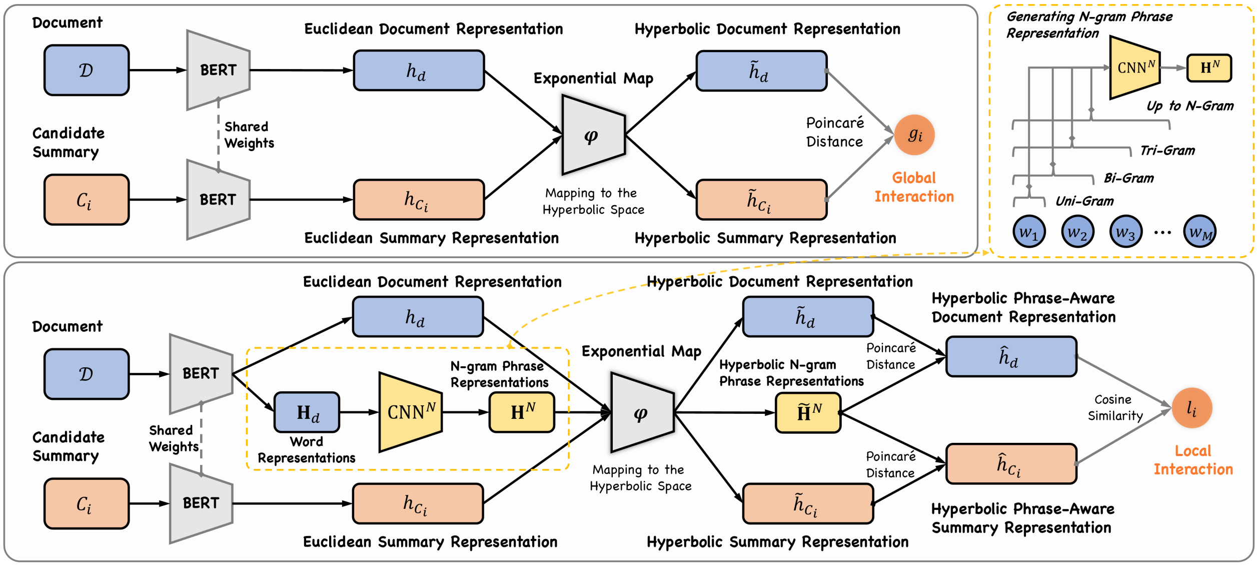

Document
$\mathcal{D}$
BERT
Shared Weights
Candidate Summary
C_i
BERT
Euclidean Document Representation
h_d
Exponential Map
φ
Mapping to the Hyperbolic Space
h_{C_i}
Euclidean Summary Representation
Hyperbolic Document Representation
$\tilde{h}_d$
Poincaré Distance
g_i
Global Interaction
$\tilde{h}_{C_i}$
Hyperbolic Summary Representation
Generating N-gram Phrase Representation
CNN^N
$\mathbf{H}^N$
Up to N-Gram
Tri-Gram
Bi-Gram
Uni-Gram
w_1 w_2 w_3 ··· w_M
Document
$\mathcal{D}$
BERT
Shared Weights
Candidate Summary
C_i
BERT
Euclidean Document Representation
h_d
$\mathbf{H}_d$
Word Representations
CNN^N
N-gram Phrase Representations
$\mathbf{H}^N$
h_{C_i}
Euclidean Summary Representation
Exponential Map
φ
Mapping to the Hyperbolic Space
Hyperbolic Document Representation
$\tilde{h}_d$
Hyperbolic N-gram Phrase Representations
$\tilde{\mathbf{H}}^N$
$\tilde{h}_{C_i}$
Hyperbolic Summary Representation
Poincaré Distance
Poincaré Distance
Hyperbolic Phrase-Aware Document Representation
$\hat{h}_d$
$\hat{h}_{C_i}$
Hyperbolic Phrase-Aware Summary Representation
Cosine Similarity
l_i
Local Interaction

Figure 2: The overall architecture of our model HISum.

matrix-vector multiplication is defined as,

$$\mathbf{M} \otimes_c \mathbf{x} = (\frac{1}{\sqrt{c}})\tanh(\frac{\|\mathbf{Mx}\|}{\|\mathbf{x}\|}\tanh^{-1}(\|\sqrt{c}\mathbf{x}\|))\frac{\mathbf{Mx}}{\|\mathbf{Mx}\|}, \tag{2}$$

where $\mathbf{M} \otimes_c \mathbf{x} = 0$ if $\mathbf{Mx} = 0$.

Poincaré Distance. The hyperbolic distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{M}_c^n$ is defined as,

$$f_c(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}}\tanh^{-1}(\sqrt{c}\| - \mathbf{x} \oplus_c \mathbf{y}\|). \tag{3}$$

In particular, when $c = 0$, the Eq. 1 denotes the Euclidean addition of two vectors in $\mathbb{R}^n$ and Eq. 3 recovers Euclidean geometry: $\lim_{c \to 0} f_c(\mathbf{x}, \mathbf{y}) = 2\|\mathbf{x} - \mathbf{y}\|$. For an open n-dimensional unit ball, the geodesics of the Poincaré disk are then circles that are orthogonal to the boundary of the ball. See Figure 1 for an illustration. Before performing operations in the hyperbolic space, a bijective projection from $\mathbb{R}^n$ to $\mathbb{M}_c^n$ that maps Euclidean vectors to the hyperbolic space is necessary. Such a projection is termed an exponential map when mapping from Euclidean space to the Poincaré model of hyperbolic geometry. The inverse to it is called a logarithmic map [21].

Exponential and Logarithmic Maps. To perform operations in the hyperbolic space, one first must define a mapping function from $\mathbb{R}^n$ to $\mathbb{M}_c^n$ to map Euclidean vectors to the hyperbolic space. Let $T_{\mathbf{x}}\mathbb{M}_c^n$ denote the *tangent space* of $\mathbb{M}_c^n$ at $\mathbf{x}$. The *exponential map* $\exp_{\mathbf{x}}^c(\cdot) : T_{\mathbf{x}}\mathbb{M}_c^n \to \mathbb{M}_c^n$ for $\mathbf{v} \neq 0$ is defined as:

$$\exp_{\mathbf{x}}^c(\mathbf{v}) = \mathbf{x} \oplus_c (\tanh(\sqrt{c}\frac{\lambda_{\mathbf{x}}^c\|\mathbf{v}\|}{2})\frac{\mathbf{v}}{\sqrt{c}\|\mathbf{v}\|}). \tag{4}$$

As the inverse of $\exp_{\mathbf{x}}^c(\cdot)$, the *logarithmic map* $\log_{\mathbf{x}}^c(\cdot) : \mathbb{M}_c^n \to T_x\mathbb{M}_c^n$ for $\mathbf{y} \neq \mathbf{x}$ is defined as:

$$\log_{\mathbf{x}}^c(\mathbf{y}) = \frac{2}{\sqrt{c}\lambda_{\mathbf{x}}^c}\tanh^{-1}(\sqrt{c}\| - \mathbf{x} \oplus_c \mathbf{y}\|)\frac{-\mathbf{x} \oplus_c \mathbf{y}}{\| - \mathbf{x} \oplus_c \mathbf{y}\|} \tag{5}$$

In practice, we follow the setting of [21, 38, 39] with the base point $x = 0$ so that the formulas are less cumbersome and empirically have little impact on the obtained results.

3 METHODOLOGY

In this section, we first introduce the problem definition of the matching-based extractive summarization task. Then, we present our hyperbolic interaction model HISum (Section 3.1), as illustrated in Figure 2, which jointly models multiple granularity interactions from a global view (Section 3.2) and a local view (Section 3.3) in the same hyperbolic space to obtain the importance scores of candidate summaries (Section 3.4). Finally, the scores are used to rank and extract the best summary (Section 3.5).

3.1 Problem Definition

As mentioned before, MatchSum [46] proposes to globally extract a summary from a summary level by a document-summary matching framework rather than from a sentence level. Specifically, MatchSum first constructs candidate summaries $C = \{C_1, C_2, ..., C_{|C|}\}$ for the document $\mathcal{D}$. Then, adopting a Siamese-BERT network to estimate the importance of each candidate summary to select the best candidate summary. In this paper, we follow the previous studies [39, 46] and adopt the same way (please refer to Zhong et al. [46] for details) to construct a set of candidate summaries C for the document $\mathcal{D}$. Given the input document $\mathcal{D} = \{w_1, ..., w_m, ..., w_M\}$ and its corresponding candidate summaries $C = \{C_1, C_2, ..., C_{|C|}\}$, a matching-based extractive summarization system aims to learn document and summary representations and model interactions between candidate summaries and their corresponding document in a semantic space to extract the best candidate summary C_{best} as the extracted summary.

3.2 Global Interaction in Hyperbolic Space

Inspired by the previous studies [3, 39, 46], we propose a Hyperbolic Siamese-BERT architecture to match the document $\mathcal{D}$ and the candidate summary C. Our Hyperbolic Siamese-BERT consists of two BERTs with shared weights (Section 3.2.1) and an interaction layer equipped with the Poincaré Distance (Section 3.2.2) to obtain the importance scores of candidate summaries.

3.2.1 Hyperbolic Siamese-BERT Encoder. To model the interaction between candidate summaries and their corresponding document from a global view, we obtain summaries and document representations by encoding the entire content of each other. Inspired by siamese network structure [3, 46], we construct a Hyperbolic Siamese-BERT encoder to represent the input document $\mathcal{D}$ and its associated candidate summary C. Specifically, each word in the document $\mathcal{D}$ is encoded by BERT as,

$$\begin{aligned} \mathbf{H} &= [h_{[\mathrm{CLS}]}^\top, h_1^\top, ..., h_m^\top, ..., h_M^\top, h_{[\mathrm{SEP}]}^\top]^\top \\ &= \mathrm{BERT}(w_i), i = 1, ..., m, ...M \ , \end{aligned} \tag{6}$$

where $h_{[\mathrm{CLS}]}$ and $h_{[\mathrm{SEP}]}$ indicate the special token of BERT. Concretely, we leverage $\mathbf{H}_d = [h_1^\top, ..., h_m^\top, ..., h_M^\top]^\top$ as word representations of the document $\mathcal{D}$ by removing the special tokens ($h_{[\mathrm{CLS}]}$ and $h_{[\mathrm{SEP}]}$) from $\mathbf{H}$. Here, we adopt $h_{[\mathrm{CLS}]}$ as the document representation h_d in our model. Later, to verify the effectiveness of different document representations, we also try to use average and max pooling strategies for $\mathbf{H}_d$ to obtain the document presentation. Furthermore, we use the same way to obtain the ground-truth summary representation h_s and the i-th candidate summary representation h_{C_i} (Here, i indicate the index of i-th candidate summary generated from the input document).

After obtaining summary and document representations, we adopt two linear transformations $\mathbf{W}_d, \mathbf{W}_c : \mathbb{R}^{d_e} \longrightarrow \mathbb{R}^{d_h}$, d_e being the dimension of contextualized embeddings in the Euclidean space and d_h being the dimension of hyperbolic space, that projects the distributed representations to the tangent space. Then the exponential map projects the tangent space to the hyperbolic space as follows,

$$\tilde{h}_d = \exp_{\mathbf{0}}^c(\mathbf{W}_d h_d), \ \tilde{h}_{C_i} = \exp_{\mathbf{0}}^c(\mathbf{W}_c h_{C_i}), \tag{7}$$

where $\tilde{h}_d$ is the hyperbolic document representation and $\tilde{h}_{C_i}$ is the i-th candidate summary representation. Concretely, $\mathbf{W}_d$ and $\mathbf{W}_c$ maps the original BERT embedding space in the Euclidean space to the tangent space of the origin of the Poincaré ball. Then $\exp_{\mathbf{0}}^{\mathrm{c}}(\cdot)$ maps the tangent space inside the Poincaré ball[2]. Consequently, we use the möbius matrix-vector multiplication as the linear transformation in the hyperbolic space[3].

3.2.2 Interacting from a Global View. Once the encoded summary representations and document representation are obtained, it is expected that every pair of candidate summaries and the document are embedded close to each other based on their geodesic distance if they are semantically similar. Therefore, we directly calculate the similarity between the i-th candidate summary h_{C_i} and its associated document h_d by the Poincaré distance[4] as a global view interaction score,

$$g_i = -f_c(\tilde{h}_{C_i}, \tilde{h}_d) \tag{8}$$

where f_c denotes the Poincaré distance and g_i indicates the global semantic similarity between the i-th candidate summary C_i and its corresponding document $\mathcal{D}$.

3.3 Local Interaction in Hyperbolic Space

Text summarization models are commonly evaluated with the standard ROUGE metric [23], reporting the F1 scores for ROUGE- 1, ROUGE-2, and ROUGE-L. These measures respectively quantify the word overlap, bigram overlap, and longest common sequence between the predicted summary and the reference summary[5]. Previous studies [30, 32] have trained extractive summarization models using a reinforcement learning objective that optimizes the mean F1 of ROUGE-1, ROUGE-2, and ROUGE-L evaluation metrics. By using the mean F1 of ROUGE-1, ROUGE-2, and ROUGE-L as the reward function, these models implicitly focus on local contextual overlaps when extracting summaries, leading to significant performance improvements. Given the findings of these studies, it is worth considering local contextual information to estimate the importance scores of candidate summaries in matching-based extractive summarization. To address this issue, we propose to capture multiple granularities of context overlaps between the document and its candidate summaries by modeling interactions from a local view. To do this, we first use a set of hyperbolic CNN encoders to obtain n-gram phrase representations in the hyperbolic space (Section 3.3.1). We then construct phrase-aware document and candidate summary representations to capture the fine-grained interactions (Section 3.3.2), which allows our model to estimate the importance of each candidate summary more accurately by considering information from context adequately.

3.3.1 Hyperbolic CNN Encoders. We adopt a set of Hyperbolic CNN encoders to obtain n-gram phrase representations in the hyperbolic space, which captures multiple granularities context of the input document. Then, the n-gram phrase representations can be obtained by the following equation,

$$\mathbf{H}^n = \mathrm{CNN}^n(\mathbf{H}_d) \tag{9}$$

where $\mathbf{H}^n$ indicates the n-gram phrase representations. Concretely, $n \in [1, N]$ is the length of phrases, and N indicates the maximum length. Specifically, each n-gram phrase has its own set of convolution filters CNN^n with window size n and stride 1. For example, when N is set to 2, we first adopt $\mathrm{CNN}^{n=1}$ and $\mathrm{CNN}^{n=2}$ to obtain n-gram phrase representations. And then, we concatenate the output embeddings by the CNN^n with different n as the final n-gram phrase representations $\mathbf{H}^N$. Similar to the equation 7, we transfer the n-gram phrase representations to the same hyperbolic space by using $\mathbf{W}_p$ and $\exp_{\mathbf{0}}^c$ to obtain hyperbolic n-gram phrase representations $\tilde{\mathbf{H}}^N$.

3.3.2 Interacting from a Local View. Here, we obtain phrase-aware document representation and the i-th phrase-aware candidate summary representation in the hyperbolic space via the Poincaré distance as follows,

$$\hat{h}_d = \frac{1}{f_c(\tilde{\mathbf{H}}^N, \tilde{h}_d)}, \hat{h}_{C_i} = \frac{1}{f_c(\tilde{\mathbf{H}}^N, \tilde{h}_{C_i})}, \tag{10}$$

[2] The choice of tangent space at the origin, instead of other points, follows previous works [14, 27] for its mathematical simplexity and optimization convenience.

[3] This transformation is theoretically redundant, and we use it primarily for numerical stability during optimization. We further note that such optimization stability is still an open problem in hyperbolic deep learning [27]. Here, we leave a detailed investigation to future work.

[4] Note that cosine similarity [44] is not appropriate to be the metric since there does not exist a clear hyperbolic inner-product for the Poincaré ball [40], so the geodesic distance is more intuitively suitable.

[5] Here, word and bigram overlap (ROUGE-1 and ROUGE-2) evaluate informativeness, whereas the longest common subsequence (ROUGE-L) determines fluency [30].

Benchmark Datasets	Domain	Type	# Data Pairs			# Input Tokens		# Extract Sentences
			Train	Valid	Test	Doc.	Sum.	
Multi-News	News Article	MDS	44, 972	5, 622	5, 622	487.3	262.0	9
PubMed	Scientific Paper	SDS	83, 233	4, 946	5, 025	444.0	209.5	6
WikiHow	Knowledge Base	SDS	168, 126	6, 000	6, 000	580.8	62.6	4

Table 1: Statistics of several benchmarks. Specifically, SDS represents single-document summarization and MDS represents multi-document summarization. The data in Doc. and Sum. denotes the average length of the document and summary in the test set, respectively. # Extract Sentences indicates the number of sentences that should extract from different datasets.

where $\hat{h}_d$ is the hyperbolic phrase-aware document representation and $\hat{h}_{C_i}$ is the hyperbolic phrase-aware representation of the i-th candidate summary. Then, we model the local interactions between the document $\hat{h}_d$ and the i-th candidate summary $\hat{h}_{C_i}$ as follows,

$$l_i = \text{cosine}(\hat{h}_d, \hat{h}_{C_i}) \tag{11}$$

where l_i indicates the local interaction score between the i-th candidate summary and its corresponding document.

3.4 Aggregating Global and Local Interactions

After obtaining interactions from both global and local views, we aggregate them as the importance of each candidate summary. To consider global and local interactions simultaneously, we simply combine g_i and l_i of the i-th candidate summary C_i together with multiplication to obtain the final importance score by,

$$s(\mathcal{D}, C_i) = g_i \cdot l_i, \tag{12}$$

where $s(\mathcal{D}, C_i)$ indicates the final importance score of the i-th candidate summary.

3.5 Training and Inference Phase

The straightforward idea is that a good candidate summary should be more semantically similar to the document than the unqualified candidate summaries [46]. Motivated by the previous studies [39, 46], we adopt a hyperbolic margin-based triplet loss to optimize the parameters:

$$\mathcal{L}_1 = \max(0, s(\mathcal{D}, C) - s(\mathcal{D}, C^*) + \delta_1) \tag{13}$$

where C indicates the candidate summary in the document $\mathcal{D}$ and δ_1 is a margin value. Through the above optimization objective, the ground-truth summary C^* should be semantically closest to the document $\mathcal{D}$ in the semantic space. Furthermore, another criterion for designing our loss function is that the candidate pair with a higher ranking disparity should have a wider margin [46]. Therefore, we adopt a hyperbolic pairwise margin-based triplet loss for ranking each candidate summary,

$$\mathcal{L}_2 = \max(0, s(\mathcal{D}, C_j) - s(\mathcal{D}, C_i) + (j - i) * \delta_2) \ (i < j), \tag{14}$$

where C_i denotes the i-th candidate summary ranked i, and δ_2 is a hyper-parameter used to distinguish between good and bad candidate summaries. Specifically, we arrange all candidate summaries in descending order of ROUGE scores with the ground-truth summary in the training phase. Finally, we combine the above two loss functions as follows,

$$\mathcal{L} = \gamma \mathcal{L}_1 + (1 - \gamma) \mathcal{L}_2. \tag{15}$$

In the inference phase, similar to Song et al. [39], Zhong et al. [46], we formulate extractive multi-document summarization as a task to search for the best summary C_{best} among all the candidates C extracted from the document $\mathcal{D}$, $C_{best} = \underset{C \in C}{\arg\max}\, s(\mathcal{D}, C)$.

4 EXPERIMENTS

4.1 Benchmark Dataset

To thoroughly verify the performance of our model HISum on documents with complex and simple latent structures, we selected both multi-document (Multi-News) and single-document (PubMed and WikiHow) datasets from various domains (e.g., news articles, scientific literature, and knowledge base). A detailed description of these benchmark datasets is illustrated in Table 1.

Multi-News dataset [13] consists of news articles and human-written summaries, which is the first large-scale multi-document summarization news dataset and comes from a diverse set of news sources. This paper uses the truncated version and concatenates the documents as a single input in all experiments.

PubMed dataset [9] is collected from scientific articles and thus consists of long documents. Similar to the previous work [46], we modify this dataset by using the introduction section as the document and the abstract section as the corresponding summary.

WikiHow dataset [20] is a diverse text corpus extracted from the online WikiHow knowledge base. Concretely, documents in this dataset span a wide range of topics.

4.2 Evaluation Metric and Baseline

Following the previous studies [7, 45, 46], we use ROUGE [23] to evaluate the extracted summary in our experiments. Concretely, ROUGE evaluates the quality of a system summary by computing overlapping lexical units. The most commonly reported metrics are ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common subsequence, LCS).

To verify the effectiveness of our model, we selected three categories of baselines, including unsupervised [4, 12, 28, 42], sentence level [13, 15, 22, 24, 43, 45], and summary level [39, 46] extractive summarization methods.

4.3 Implementation Detail

In this paper, we adopt the "bert-base-uncased" version to initialize the hyperbolic Siamese-BERT encoder as the backbone. Since the maximum position in the pre-trained position embedding of BERT is 512, we truncate the documents to 512 words. The embedding

Extractive Summarization Model	Multi-News Dataset			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-AVERAGE
MATCH-ORACLE [46]	47.45	17.41	43.14	36.00
Unsupervised Extractive Summarization Model				
LEAD [36]	43.08	14.27	38.97	32.11
TextRank† [28]	41.95	13.86	38.07	31.29
LexRank† [12]	41.77	13.81	37.87	31.15
MMR† [4]	44.72	14.92	40.07	33.24
Supervised Sentence Level Extractive Summarization Model				
PG† [22]	44.55	15.54	40.75	33.61
BottomUp† [15]	45.27	15.32	41.38	33.99
Hi-MAP† [13]	45.21	16.29	41.39	34.30
HDSG† [43]	46.05	16.35	42.08	34.83
PRESUMM [45]	46.34	16.88	42.20	35.14
Supervised Summary Level Extractive Summarization Model				
MatchSum [46]	46.20	16.51	41.89	34.87
HyperSiameseNet [39]	46.67	16.93	42.38	35.33
HISum	**47.14**	**17.24**	**42.84**	**35.74**

Table 2: Performance on the Multi-News test set. The models with † indicates that the results provided by Wang et al. [43]. The results of baselines are those presented in the original papers or better results published in other papers recently.

Model	PubMed Dataset		
	ROUGE-1	ROUGE-2	ROUGE-L
MATCH-ORACLE [46]	42.21	15.42	37.67
LEAD [36]	35.63	12.28	25.17
BERTEXT [24]	41.05	14.88	36.57
MatchSum [46]	41.21	14.91	36.75
HISum	**41.64**	**15.13**	**37.14**

Table 3: Performance on the PubMed test set.

dimension d_e is 768, the same as the other embeddings in the "BERT-base" model. A dropout layer is added after the hyperbolic CNN encoders with a dropout probability set to 0.1. Specifically, $\gamma = 0.5$, $d_e = 768$, and $d_h = 512$. To prevent performance degradation, we choose $\delta_1 = 0.01$ and $\delta_2 = 0$.

In the same way that gradient-based optimization methods are used for trainable parameters in the Euclidean space, the hyperbolic parameters can be updated via Riemannian adaptive optimization methods [2]. Therefore, similar to the previous work [39], we use the Riemannian Adam [2] optimizer with warming-up is used. The learning rate schedule follows [41] as $lr = 2e^{-3} \cdot \min(step^{-0.5}, step \cdot wm^{-1.5})$, where each step is a batch size of 16 and wm indicates warm-up steps of 10,000. Our model is trained with 8 A4000-16G GPUs. The mini-batch size is set to 16 to utilize the GPU memory fully. In addition, we also record the performance of the best checkpoints on the test set by keeping three top checkpoints on the validation set during training.

4.4 Overall Performance

As illustrated in Table 2, we present our main results on the Multi-News dataset. It includes the automatic evaluation results using the ROUGE-F1 metric. As shown in Table 2, our model HISum outperforms the unsupervised extractive summarization models such as LEAD [36], TextRank [28], LexRank [12], and MMR [4] by a large margin. This is not surprising since the supervised methods are trained end-to-end with supervised data. Compared to the supervised sentence level models (e.g., PRESUMM [45]) whose sentences are extracted individually, HISum achieves superior performance. This is probably because it still relies on the pointer-based network or sequence labeling methods, which select sentences one by one rather than distinguishing the semantics of different summaries from a global view. Furthermore, compared to the supervised summary level models MatchSum [46] and HyperSiameseNet [39], our model HISum also achieves better performance on the Multi-News datasets, indicating that capturing global and local interactions simultaneously in the hyperbolic space for extractive multi-document summarization is an effective and valuable strategy.

Generally, the input documents containing multiple documents have more complex latent hierarchy information than the input documents containing single documents. Therefore, we conduct our model on two benchmark single-document summarization datasets (e.g., WikiHow and PubMed) in Table 3 and Table 4. Compared to the supervised summary level baseline MatchSum, our model HISum achieves significant improvements on the two datasets. These results further demonstrate the effectiveness of our hyperbolic interaction model HISum and show the potential of the hyperbolic space in the extractive summarization task.

4.5 Ablation Test

To study the contribution of each component in our model HISum. We remove each component and report the results in Table 4. Concretely, HISum *w/o global-view interaction* indicates our model HISum only models interaction from a local view for estimating the importance of each candidate summary. Overall, we can see that jointly modeling interactions from global and local views are helpful in improving the performance of extractive summarization. Meanwhile, the performance of HISum drops when removing them.

Model	**WikiHow Dataset**		
	ROUGE-1	**ROUGE-2**	**ROUGE-L**
MATCH-ORACLE [46]	35.22	10.55	32.87
LEAD [36]	24.97	5.83	23.24
TextRank [28]	21.64	5.34	19.68
LexRank [12]	25.46	5.89	23.63
MMR [4]	22.02	4.40	20.22
BERTEXT [24]	30.31	8.71	28.24
MatchSum [46]	31.85	8.98	29.58
HISum	**32.80**	**9.47**	**30.53**
HISum *w/o local interaction*	32.36	9.14	30.06
HISum *w/o global interaction*	32.02	8.96	29.76

Table 4: Performance on the WikiHow test set.

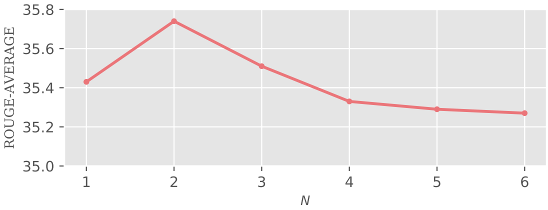


Figure 3: Effect of hyperbolic phrase-aware document representation with different N on the Multi-News dataset.

The results of our model HISum, whether only modeling interaction from a local view (HISum w/o local-view interaction) or only capturing interaction from a global view (HISum w/o global-view interaction), are significantly better than MatchSum, which further proves the effectiveness of capturing extractive summarization model in hyperbolic space.

4.6 Effect of Different Granularities (N)

To further study the effectiveness of capturing interactions from a local view, we verify different granularities N of n-gram phrase-aware document and candidate summary representations. Here, we present the results in Figure 3. When N is set to 2, it collects both uni-gram and bi-gram phrase representations of the source document to construct the phrase-aware document and candidate summary representations. As illustrated in Figure 3, we find HISum achieves the best results when N is set to 2.

4.7 Effect of Different Dimensions (d_h)

Generally, when embedding in the hyperbolic space, adopting a lower dimension for the embedding space usually get better performance for the downstream NLP tasks. Therefore, we report the results of the hyperbolic document and candidate summary representations with different dimensions d_h, as shown in Table 4. The results show that obtaining hyperbolic representations by adopting a smaller dimension is not appropriate for our model, suffering from poor performance. Instead, a slightly larger value (i.e., 512) yields the best performance. We consider the main reason may be that our model uses the pre-trained language model (i.e., BERT) to obtain initial embeddings in the Euclidean space and maps them to the hyperbolic space rather than directly using representations learned from the hyperbolic space. Therefore, if the dimension of hyperbolic space is reduced too much when projecting representations from the Euclidean space to the hyperbolic space (e.g., 768 ⟶ 64), damaging the representations obtained by BERT, where we consider 512 may be a trade-off value for the above issue.

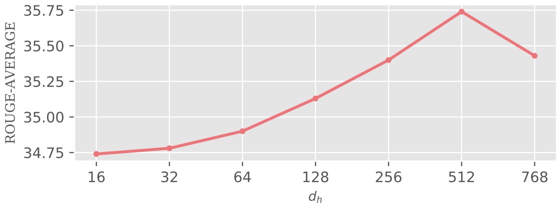


Figure 4: Effect of hyperbolic representation with different dimensions d_h on the Multi-News dataset.

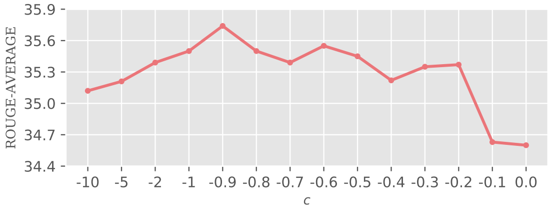


Figure 5: Effect of different curvatures of the Poincaré ball. As the curvature goes closer to 0, the Poincaré ball behaves more similarly to euclidean space.

4.8 Effect of Different Curvatures (c)

We further investigate the property of the hyperbolic space with different curvatures, which measures how "curved" the space is. Additionally, if we gradually change the curvature to 0, the space would be "less curved" and more similar to the Euclidean space (more "flat"). Consequently, the Poincaré scores converge to the Euclidean scores. When the curvature is 0, we recover the Euclidean space. As illustrated in Figure 5, the optimal curvature is -0.9.

4.9 Effect of Different Hyperbolic Document and Summary Representations

In our model, we empirically adopt the "[CLS]" token of the Hyperbolic Siamese-BERT encoder as the document representation and employ the max-pooling operation to obtain the representations of candidate summaries via the Hyperbolic Siamese-BERT encoder.

As illustrated in Table 5, we also present the comparison of different document and candidate summary representations, which are obtained by several methods: the "[CLS]" token, average-pooling, and max-pooling. Specifically, we employ the max-pooling and average-pooling functions to get document representation from the output of the last BERT layer. From the results, we can see that the "[CLS]" token is the best choice for document representation,

Document Embedding	Summary Embedding	Multi-News Dataset			
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-AVERAGE
CLS Token	CLS Token	46.89	17.07	42.59	35.52
	Average Pooling	46.52	16.83	42.21	35.19
	Max Pooling	**47.14**	**17.24**	**42.84**	**35.74**
Average Pooling	CLS Token	46.91	17.10	42.61	35.54
	Average Pooling	46.53	16.85	42.24	35.21
	Max Pooling	46.46	16.82	42.18	35.15
Max Pooling	CLS Token	46.86	17.06	42.56	35.49
	Average Pooling	46.50	16.86	42.21	35.19
	Max Pooling	46.81	17.05	42.52	35.46

Table 5: Effect of different document and candidate summary embeddings. Average Pooling and Max Pooling strategies are employed on the output of the last BERT layer to produce document and candidate summary embeddings.

and the max-pooling operation is the best choice for representation candidate summaries in our model.

5 RELATED WORK

5.1 Extractive Multi-Document Summarization

Document summarization is a crucial sub-task of natural language processing and is usually divided into two categories: extractive and abstractive. An extractive summarization is a combination of sentences extracted from the original document. These sentences are calculated to carry the salient content of their corresponding document. In contrast, abstractive summarization is a technique that examines the document and produces a new text. In other words, this method creates novel sentences based on the most critical information of the originals. Even though people usually use abstract ways to create a summary, extractive summarization methods focus more on recent studies [1]. The performances of extractive summarizing systems are often better than abstractive summarizing systems [12].

Extractive summarization has been widely studied, and it can be categorized from the following perspective: single vs. multi-document. A notable challenge for Multi-Document Summarization (MDS) compared to single-document summaries is the extremely-long length of the input, suffering from complex latent hierarchical structures. Specifically, extractive multi-document summarization methods attempt to create a summary by selecting a set of sentences from several relevant documents that are most salient and relevant. To tackle this challenging task, various techniques and methods [39, 45, 46] have been applied in extractive multi-document summarization.

With the development of pre-trained language models [11, 26], the paradigms of many Natural Language Processing (NLP) downstream tasks have generated more changes. More and more work focuses on exploring a new paradigm for different downstream NLP tasks. Therefore, recent studies have attempted to build two-stage document summarization systems [7, 46]. Specific to extractive summarization, the first stage usually extracts some fragments of the original text, and the second stage further selects or modifies them based on these fragments. Specifically, the second stage first obtains representations by pre-trained language models and then extracts the summary by calculating the semantic similarity.

Different from the existing matching-based extractive summarization models, we propose a new hyperbolic interaction model to jointly model interactions from global and local views for capturing sufficient context information and further explore extractive summarization methods in hyperbolic space.

5.2 Hyperbolic Representation Learning

Hyperbolic geometry has been widely investigated in representation learning in recent years due to its excellent expression capacity in capturing complex data with non-Euclidean properties [17, 34, 35, 48]. To construct neural networks in the hyperbolic space, previous studies combine the formalism of Möbius gyrovector spaces with the Riemannian geometry, derive hyperbolic versions of critical mathematical operations such as Möbius matrix-vector multiplication, and leverage them to build hyperbolic neural networks [14]. Recently, many methods using hyperbolic geometry have been proposed for various downstream tasks due to its better inductive bias for capturing latent hierarchical information than Euclidean space. Specifically, in natural language processing, hyperbolic representation learning has been successfully applied to capturing latent hierarchical structures for learning context embeddings [5, 10, 40]. For example, Poincaré embeddings [31] and Poincaré Glove [40] learn text embeddings of hierarchies using Poincaré models and exhibit impressive results.

6 CONCLUSION AND FUTURE WORK

In this paper, we point out that most existing matching-based extractive summarization models ignore the local interaction between the document and candidate summaries and propose a novel hyperbolic interaction model for extractive multi-document summarization, which jointly models global and local interactions to estimate the importance score of each candidate summary for extracting the best candidate summary. Extensive experiment results on three public benchmarks demonstrate that our model can effectively capture global and local information and achieve remarkable results.

In future work, it would be interesting and valuable to model more interactions between the document and its corresponding candidate summaries from multiple perspectives by mining and introducing latent knowledge (e.g., topics or entities or keyphrases) of the source document.

ACKNOWLEDGMENTS

We thank the three anonymous reviewers for carefully reading our paper and their insightful comments and suggestions. This work was partly supported by the Fundamental Research Funds for the Central Universities (2019JBZ110); the National Natural Science Foundation of China under Grant 62176020; the National Key Research and Development Program (2020AAA0106800); the Beijing Natural Science Foundation under Grant L211016; CAAI-Huawei MindSpore Open Fund; and Chinese Academy of Sciences (OEIP-O-202004).

REFERENCES

[1] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys J. Kochut. 2017. Text Summarization Techniques: A Brief Survey. *CoRR* abs/1707.02268 (2017). arXiv:1707.02268 http://arxiv.org/abs/1707.02268

[2] Gary Bécigneul and Octavian-Eugen Ganea. 2019. Riemannian Adaptive Optimization Methods. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=r1eiqi09K7

[3] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature Verification Using A "Siamese" Time Delay Neural Network. *Int. J. Pattern Recognit. Artif. Intell.* 7, 4 (1993), 669–688. https://doi.org/10.1142/S0218001493000339

[4] Jaime G. Carbonell and Jade Goldstein. 2017. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *SIGIR Forum* 51, 2 (2017), 209–210. https://doi.org/10.1145/3130348.3130369

[5] Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing BERT in Hyperbolic Spaces. In *International Conference on Learning Representations*. https://openreview.net/forum?id=17VnwXYZyhH

[6] Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. 2020. Hyperbolic Interaction Model for Hierarchical Multi-Label Classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 7496–7503. https://ojs.aaai.org/index.php/AAAI/article/view/6247

[7] Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 675–686. https://doi.org/10.18653/v1/P18-1063

[8] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words.. In *ACL (1)*. The Association for Computer Linguistics. http://dblp.uni-trier.de/db/conf/acl/acl2016-1.html#0001L16

[9] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 615–621. https://doi.org/10.18653/v1/n18-2097

[10] Shuyang Dai, Zhe Gan, Yu Cheng, Chenyang Tao, Lawrence Carin, and Jingjing Liu. 2020. APo-VAE: Text Generation in Hyperbolic Space. *CoRR* abs/2005.00054 (2020). http://dblp.uni-trier.de/db/journals/corr/corr2005.html#abs-2005-00054

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. In *NAACL-HLT (1)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19

[12] Günes Erkan and Dragomir R. Radev. 2011. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *CoRR* abs/1109.2128 (2011). arXiv:1109.2128 http://arxiv.org/abs/1109.2128

[13] Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1074–1084. https://doi.org/10.18653/v1/p19-1102

[14] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic Neural Networks.. In *NeurIPS*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 5350–5360. http://dblp.uni-trier.de/db/conf/nips/nips2018.html#GaneaBH18

[15] Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4098–4109. https://doi.org/10.18653/v1/d18-1443

[16] Min Gui, Zhengkun Zhang, Zhenglu Yang, Yanhui Gu, and Guandong Xu. 2018. An Effective Joint Framework for Document Summarization. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 121–122. https://doi.org/10.1145/3184558.3186959

[17] Matthias Hamann. 2018. On the tree-likeness of hyperbolic spaces. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 164. 345–361.

[18] C. Hopper and B. Andrews. 2011. *The Ricci Flow in Riemannian Geometry.* The Ricci flow in Riemannian geometry.

[19] Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss.. In *ACL (1)*. Association for Computational Linguistics, 132–141. http://dblp.uni-trier.de/db/conf/acl/acl2018-1.html#SunHLLMT18

[20] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1967–1972. https://doi.org/10.18653/v1/d15-1229

[21] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan V. Oseledets, and Victor S. Lempitsky. 2020. Hyperbolic Image Embeddings.. In *CVPR*. IEEE, 6417–6427. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2020.html#KhrulkovMUOL20

[22] Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4131–4141. https://doi.org/10.18653/v1/d18-1446

[23] C. Y. Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*. Barcelona, Spain.

[24] Yang Liu. 2019. Fine-tune BERT for Extractive Summarization. *CoRR* abs/1903.10318 (2019). http://dblp.uni-trier.de/db/journals/corr/corr1903.html#abs-1903-10318

[25] Yizhu Liu, Qi Jia, and Kenny Q. Zhu. 2021. Keyword-aware Abstractive Summarization by Extracting Set-level Intermediate Summaries. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 3042–3054. https://doi.org/10.1145/3442381.3449906

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). http://dblp.uni-trier.de/db/journals/corr/corr1907.html#abs-1907-11692

[27] Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. 2019. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 12544–12555. https://proceedings.neurips.cc/paper/2019/hash/0ec04cb3912c4f08874dd03716f80df1-Abstract.html

[28] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text.. In *EMNLP*. ACL, 404–411. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2004.html#MihalceaT04

[29] Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016. Classify or Select: Neural Architectures for Extractive Document Summarization. *CoRR* abs/1611.04244 (2016). http://dblp.uni-trier.de/db/journals/corr/corr1611.html#NallapatiZM16

[30] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking Sentences for Extractive Summarization with Reinforcement Learning.. In *NAACL-HLT*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 1747–1759. http://dblp.uni-trier.de/db/conf/naacl/naacl2018-1.html#NarayanCL18

[31] Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations.. In *NIPS*. 6338–6347. http://dblp.uni-trier.de/db/conf/nips/nips2017.html#NickelK17

[32] Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A Deep Reinforced Model for Abstractive Summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018,*

Conference Track Proceedings. OpenReview.net. https://openreview.net/forum?id=HkAClQgA-

[33] Haggai Roitman, Guy Feigenblat, Doron Cohen, Odellia Boni, and David Konopnicki. 2020. Unsupervised Dual-Cascade Learning with Pseudo-Feedback Distillation for Query-Focused Extractive Summarization. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 2577–2584. https://doi.org/10.1145/3366423.3380009

[34] Christopher De Sa, Albert Gu, Christopher Ré, and Frederic Sala. 2018. Representation Tradeoffs for Hyperbolic Embeddings. *CoRR* abs/1804.03329 (2018). arXiv:1804.03329 http://arxiv.org/abs/1804.03329

[35] Rik Sarkar. 2011. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane.. In *Graph Drawing (Lecture Notes in Computer Science, Vol. 7034)*, Marc J. van Kreveld and Bettina Speckmann (Eds.). Springer, 355–366. http://dblp.uni-trier.de/db/conf/gd/gd2011.html#Sarkar11

[36] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks.. In *ACL (1)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 1073–1083. http://dblp.uni-trier.de/db/conf/acl/acl2017-1.html#SeeLM17

[37] Jiaxin Shi, Chen Liang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Hanwang Zhang. 2019. DeepChannel: Salience Estimation by Contrastive Learning for Extractive Document Summarization.. In *AAAI*. AAAI Press, 6999–7006. http://dblp.uni-trier.de/db/conf/aaai/aaai2019.html#ShiLHL0Z19

[38] Mingyang Song, Yi Feng, and Liping Jing. 2022. Hyperbolic Relevance Matching for Neural Keyphrase Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 5710–5720. https://doi.org/10.18653/v1/2022.naacl-main.419

[39] Mingyang Song, Yi Feng, and Liping Jing. 2022. A Preliminary Exploration of Extractive Multi-Document Summarization in Hyperbolic Space. In *Proceedings of the 31st ACM International Conference on Information &amp; Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4505–4509. https://doi.org/10.1145/3511808.3557538

[40] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2019. Poincare Glove: Hyperbolic Word Embeddings.. In *ICLR (Poster)*. OpenReview.net. http://dblp.uni-trier.de/db/conf/iclr/iclr2019.html#TifreaBG19

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need.. In *NIPS*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. http://dblp.uni-trier.de/db/conf/nips/nips2017.html#VaswaniSPUJGKP17

[42] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2016. Order Matters: Sequence to sequence for sets.. In *ICLR (Poster)*, Yoshua Bengio and Yann LeCun (Eds.). http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#VinyalsBK15

[43] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization.. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 6209–6219. http://dblp.uni-trier.de/db/conf/acl/acl2020.html#WangLZQH20

[44] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 4144–4150. https://doi.org/10.24963/ijcai.2017/579

[45] Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen R. McKeown, and Snigdha Chaturvedi. 2022. Read Top News First: A Document Reordering Approach for Multi-Document News Summarization. *CoRR* abs/2203.10254 (2022). https://doi.org/10.48550/arXiv.2203.10254 arXiv:2203.10254

[46] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive Summarization as Text Matching.. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 6197–6208. http://dblp.uni-trier.de/db/conf/acl/acl2020.html#ZhongLCWQH20

[47] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences.. In *ACL (1)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 654–663. http://dblp.uni-trier.de/db/conf/acl/acl2018-1.html#ZhaoZWYHZ18

[48] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. 2020. HyperText: Endowing FastText with Hyperbolic Geometry. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1166–1171. https://doi.org/10.18653/v1/2020.findings-emnlp.104