Optimal Transport in Summarisation: Towards Unsupervised Multimodal Summarisation

PEGGY PIK-YEE TANG

Supervisor: Associate Professor Zhiyong Wang

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

30 August 2023

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature,

Peggy Pik-yee Tang

Authorship Attribution Statement

The publications associated with this thesis are as follows.

- Chapter 2 of this thesis is with the data summarised from the following publications associated with individual chapters.
- Chapter 3 of this thesis is published with:

  Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. 2022. OTExtSum: Extractive Text Summarisation with Optimal Transport. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1128–1141, Seattle, United States. Association for Computational Linguistics.

  I am the first author. I designed the algorithms, conducted the experiments and wrote the drafts.
- Chapter 4 of this thesis is published with:

  Peggy Tang, Junbin Gao, Lei Zhang, and Zhiyong Wang. 2023. Efficient and Interpretable Compressive Text Summarisation with Unsupervised Dual-Agent Reinforcement Learning. In Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 227–238, Toronto, Canada. Association for Computational Linguistics.

  I am the first author. I designed the algorithms, conducted the experiments and wrote the drafts.
- Chapter 5 of this thesis is published with:

  Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. TLDW: Extreme Multimodal Summarisation of News Videos. In IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). (In press)

  I am the first author. I designed the algorithms, conducted the experiments and wrote the drafts.
- Chapter 6 of this thesis is published with:

Peggy Tang, Kun Hu, Lei Zhang, Junbin Gao, Jiebo Luo, and Zhiyong Wang. 2023. TopicCAT: Unsupervised Topic-Guided Co-Attention Transformer for Extreme Multimodal Summarisation. In ACM International Conference on Multimedia (ACM MM). (Accepted)

I am the first author. I designed the algorithms, conducted the experiments and wrote the drafts.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Student Name: Peggy Pik-yee Tang
Signature,
Peggy Pik-yee Tang
Date: 30 August 2023

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Zhiyong Wang
Signature,
Zhiyong Wang
Date: 30 August 2023

Abstract

Summarisation aims to condense a given piece of information into a short and succinct summary that best covers its semantics with the least redundancy. This helps users quickly browse and understand long content by focusing on the most important ideas. Summarisation on a single modality, such as text summarisation, has been actively studied for decades. Text summarisation is a challenging task in natural language processing that condenses a document into a succinct summary. With the explosion of multimedia data, multimodal summarisation with multimodal output emerges and extends the inquisitiveness of the task. Summarising a video-document pair into a visual-textual summary helps users obtain a more informative and visual understanding. Although various methods have achieved promising performance, they have limitations, including expensive training, lack of interpretability, and insufficient brevity.

Therefore, this thesis addresses the gap and examines the application of optimal transport (OT) in unsupervised summarisation, and the major contributions are as follows: (1) An interpretable OT-based method is proposed for text summarisation. It formulates summary sentence extraction as minimising the transportation cost to a given document regarding their semantic distributions; (2) An efficient and interpretable unsupervised reinforcement learning method is proposed for text summarisation. Multi-head attentional pointer-based networks are designed to learn the representation and extract salient sentences and words. The learning strategy aims to mimic human judgment by optimising summary quality regarding OT-based semantic coverage and fluency; (3) A new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO) is introduced. It summarises a video-document pair into an extremely short multimodal summary. An unsupervised Hierarchical Optimal Transport Network is proposed to learn across multiple modalities and use OT solvers to maximise multimodal semantic coverage. A new large-scale dataset is constructed to facilitate future research in this new direction; (4) A Topic-Guided Co-Attention Transformer method is proposed for XMSMO. It constructs a two-stage unimodal and cross-modal modelling with

clustering-based cross-modal topic guidance. A novel OT-guided unsupervised training strategy optimises from the perspective of the similarity between semantic distributions of topics. Comprehensive experiments demonstrate the effectiveness of the proposed methods.

Acknowledgements

I began pursuing a PhD in Computer Science at the University of Sydney in 2020, which has been an exciting experience for me. Despite the challenges posed by the COVID-19 pandemic, I have learned and adapted quickly, which has helped me become more creative. Throughout my research project and the completion of my thesis, I received support and assistance from numerous individuals, and I would like to express my gratitude to them for their contribution to my PhD journey.

I first and foremost thank the Lord God Almighty for His steadfast love, protection, and the devotion He placed in my heart to persevere to the end.

I am thankful for the support and guidance of my supervisor A/Prof Zhiyong Wang, who provided continuous supervision throughout my PhD study. His expertise, dedication, and support enabled me to explore an emerging research field and achieve success during my candidature. Additionally, I would like to thank my collaborators, Prof Junbin Gao, Dr Kun Hu, Prof Jiebo Luo, A/Prof Rui Yan, and Dr Lei Zhang, for their invaluable feedback and advice on my PhD research.

I am grateful for the care and support my group of friends provided throughout the years. I thank my friends in the Multimedia Computing Lab for sharing their knowledge with me. I acknowledge the opportunity to work with my fellow Peer Learning Advisors at the University of Sydney, which has connected me to the broader university communities and given me a more comprehensive understanding of how technology is used in different disciplines to enhance people's lives. I thank my friends that I met at the Luke Fellowship for all the kind support, fun time, and mental and spiritual care you have shared with me over the years.

Finally, I would like to express my gratitude to my beloved family members: my husband, Gordon; my parents, Edward and May; and my siblings, Fiona and Jimmy. They have always provided love, support, and companionship for me. I am excited about the future time we will spend together and the memories we will create.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Background and Research Motivation

Summarisation is the process of compacting information into a brief and concise summary that captures the main meaning while minimising repetition. Its purpose is to assist users in efficiently skimming through and comprehending lengthy content by highlighting key ideas. While summarisation on a single modality, such as text summarisation [77, 106, 60, 48], has been the subject of extensive research for decades, multimodal summarisation [141, 142, 52] has attracted research interest in the past few years.

Text summarisation has always been an interesting yet challenging task in the natural language processing field. It aims to transform the original document into a shorter version that covers the main ideas, i.e. a summary [70, 83]. Doing so helps reduce the time for readers to acquire knowledge and discover relevant information. It is one of the most challenging tasks in natural language processing; it involves the ability to understand the meaning of the original text and the ability to synthesise it in natural language. Even for a trained human abstractor, journalist, or professional writer, it is still an intricate craft to master and is regarded as a form of fine art [5]. This technique has broad applications in our everyday life. Figure 1.1 shows some examples of text summarisation applications. For example, the most common one is the news summarisation. Instead of reading an enormous amount of news articles, readers could grasp the highlights of events that matter at a glance. Book summaries give an overview of the story plot and the essence of the ideas to attract readers to continue reading the book. Abstracts of scientific articles highlight major findings and contributions, allowing readers to acquire knowledge quickly.

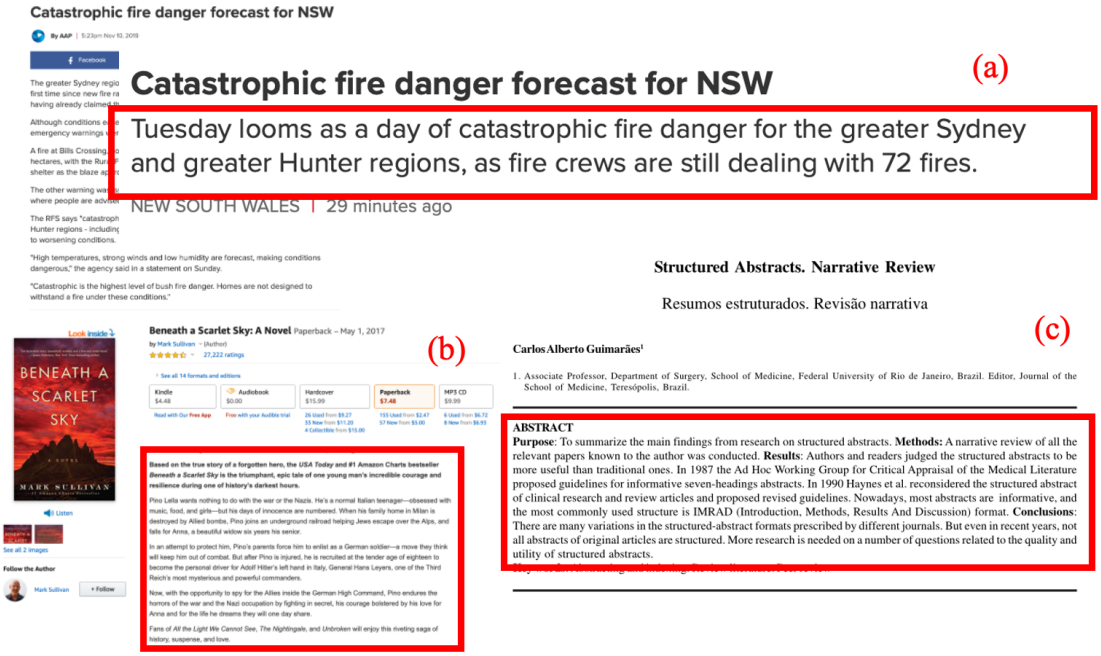

FIGURE 1.1. Some examples of text summarisation applications, including (a) news summary, (b) book summary, and (c) abstract of a scientific article.

With the exponentially increasing amount of text data and the request for instant knowledge, it becomes infeasible and costly to generate summaries manually promptly. There is a growing demand to develop and advance automatic text summarisation techniques. Early research on text summarisation started as early as the 1950s [64]. Traditional methods apply different techniques to identify the most important sentences or words in the input document and combine them to form an output summary. They rely on handcrafted features, which mainly depend on linguistic knowledge. Since more textual data are becoming available nowadays, most proposed methods are deep learning-based since the first proposed in 2015[104]. An input document is transformed into an output summary by deep learning models based on features automatically learnt from the data. There are several comprehensive surveys of this field [109][105][62][22][43].

While text summarisation has been investigated for decades, with the rapid growth of multimedia data [143], there is an emerging interest in Multimodal Summarisation with Multimodal Output (MSMO) [141, 142, 52] in recent years. MSMO aims to summarise a pair of a video or a set of images and a document into a visual-textual summary, since image and text could

complement each other, where images help users to grasp events while texts provide more details related to the events. Figures 1.2 and 1.3 show some examples of MSMO applications. Such summarisation empowers users to swiftly identify the key aspects of multimodal content, determining if further reading or watching is worthwhile according to multimodal summaries with complementary visual and text content. This help users to better obtain a more informative and visual understanding of events. Also, this can help make the information more accessible to users with various individual needs. For example, users with a reading disorder may refer more to the visual summary, and users with vision impairments may find the textual summary more accessible using a screen reader.

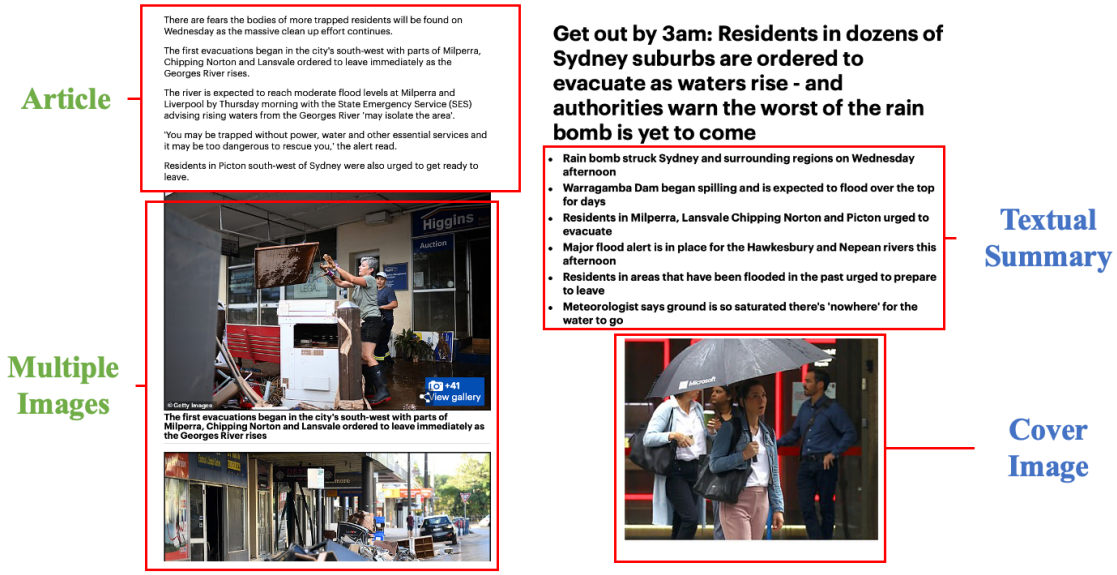

FIGURE 1.2. Example of MSMO, which summarises a set of images and a document into a visual-textual summary.

After years of efforts by the research community, these existing methods have advanced the task of text summarisation and multimodal summarisation. However, both of them remain an open problem, and the current state-of-the-art can still not approximate a human abstractor. One of the limitations is that existing text summarisation methods [125] often first score the importance of individual sentences of a given document and then combine the top-ranked ones to form a summary. However, the sentences with high importance scores may not represent the document from a global perspective [45], resulting in a sub-optimal and redundant summary. Another limitation is the expensive training cost of existing methods and

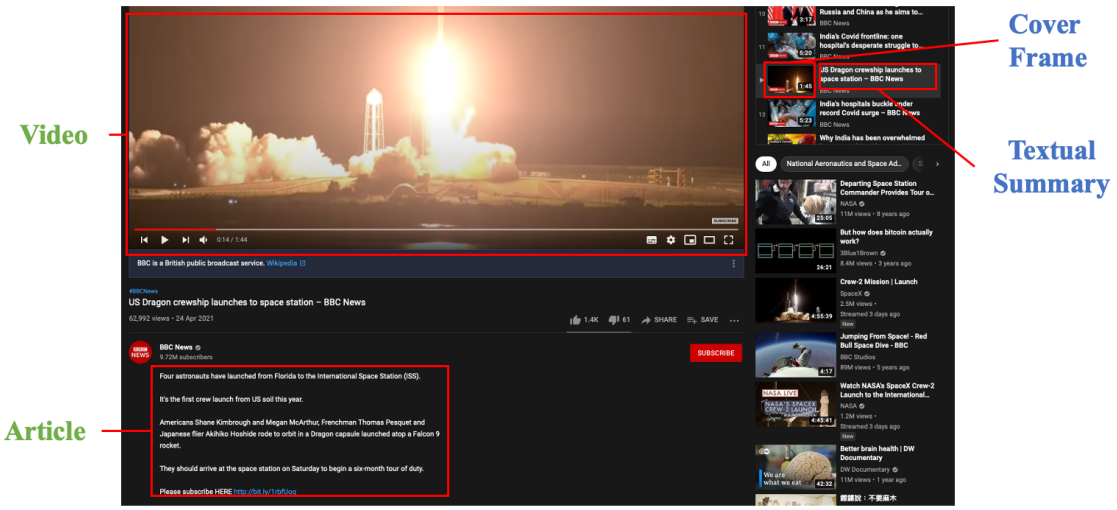

FIGURE 1.3. Example of MSMO application summarising a video and document into a visual-textual summary.

the lack of interpretability for the summarisation process. Most existing models are trained by supervised methods, which may not be able to approximate human judgment.

For MSMO, one of the limitations is that most of the existing MSMO methods are designed for short visual inputs, such as short videos and multiple images, without considering the summary length. According to the statistics of YouTube[1], one of the most popular online video-sharing platforms, in 2023, more than 500 hours of video contents are uploaded every minute; and more than a billion logged-in users visit YouTube each month and watch over a billion hours of video every day. In particular, YouTube channels of news organisations are a significant part of the content, which often have millions of subscribers and views. For example, as of May 2023, the British Broadcasting Corporation News YouTube channel had over 14 million subscribers and had accumulated over 4 billion views[2]. Given the increasing pace of producing multimedia data and the subsequent challenge of keeping up with the explosive growth of such rich content, these existing methods may be sub-optimal to address the imminent issue of information overload of multimedia data. Moreover, most existing methods are supervised, requiring the resource-expensive collection of ground-truth summaries as supervision.

[1] https://blog.youtube/press/
[2] https://www.youtube.com/@BBCNews/about

This thesis identifies four objectives to address the research gap of text summarisation and MSMO. Our first objective is to improve the global-level optimisation for summarisation by investigating the conceptualisation of text summarisation as an optimal transport problem for the first time. There have been some studies on OT in NLP, such as document distance [47, 130], text generation [10], text matching [111], and machine translation [124]. These methods generally focus on deriving similarities between words, sentences, and documents. On the contrary, we for the first time formulate text summarisation as an OT problem that optimally transports the semantic distributions between two texts (e.g., source document and summary candidate). The second objective is to improve human judgment approximation and reduce reliance on parallel training data by investigating unsupervised reinforcement-based text summarisation method. The existing supervised learning methods trained with ground-truth summaries [106, 131, 119, 61] and with reinforcement training to optimise the ROUGE metric [92, 89] may not provide useful insights on human judgment approximation . The third objective is to improve the information overload issue of multimedia data by investigating a new task of extreme multimodal summarisation and collecting a new dataset, and investigating an unsupervised hierarchical neural method for it. Most of the existing MSMO methods [141, 142, 52] are designed for short visual inputs, such as short videos and multiple images, without considering the summary length. Given the increasing pace of producing multimedia data and the subsequent challenge in keeping up with the explosive growth of such rich content, these existing methods may be sub-optimal to address the imminent issue of information overload of multimedia data. The fourth objective is to further enhance the extreme multimodal summarisation task by investigating unsupervised topic-guided neural method to capture the overarching themes in documents and videos effectively. As topic models are generally useful for identifying key aspects from long documents [80], our proposed new architecture is devised for the extreme multimodal summarisation task.

1.2 Contributions

The main contributions of this thesis are as follows:

(1) A non-learning-based extractive text summarisation method - OTExtSum is proposed by treating the text summarisation task as an optimal transport problem for the first time. Two optimisation strategies for OTExtSum are designed to optimise this problem formulation: beam search strategy and binary integer programming strategy.

(2) The first unsupervised compressive text summarisation method with dual-agent reinforcement learning, namely URLComSum, is proposed. An efficient and interpretable multi-head attentional pointer-based neural network is designed to learn the representation and extract salient sentences and words. The unsupervised reinforcement learning strategy is designed to mimic human judgment by optimising summary quality in terms of the semantic coverage reward, measured by Wasserstein distance, and the fluency reward, measured by Syntactic Log-Odds Ratio (SLOR).

(3) A new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO), is proposed. It aims to summarise a video-document pair into an extreme multimodal summary (i.e., one cover frame as the visual summary and one sentence as the textual summary). A novel unsupervised Hierarchical Optimal Transport Network (HOT-Net) is proposed. The hierarchical encoding and decoding are conducted across visual and textual modalities, and optimal transport solvers are introduced to guide the summaries to maximise their semantic coverage. A new large-scale dataset XMSMO-News is constructed for the research community to facilitate research in this new direction.

(4) A novel transformer architecture - Topic-Guided Co-Attention Transformer (TopicCAT) - is proposed for emerging extreme multimodal summarisation. It constructs a two-stage learning strategy for unimodal and cross-modal modelling with clustering-based cross-modal topic guidance. A novel optimal transport-guided unsupervised training strategy is devised to optimise TopicCAT from the perspective of the similarity between semantic distributions of topics.

1.3 Thesis Structure

The rest of the thesis is organised as follows:

Chapter 2 presents the background and the related studies for text summarisation and multimodal summarisation with multimodal output methods.

Chapter 3 presents an interpretable optimal transport-based method (OTExtSum) for text summarisation. It formulates summary sentence extraction as minimising the transportation cost to a given document regarding their semantic distributions.

Chapter 4 presents an unsupervised reinforcement learning method (URLComSum) for text summarisation. Multi-head attentional pointer-based networks learn the representation and extract salient sentences and words. The learning strategy mimics human judgment by optimising summary quality regarding semantic coverage and fluency.

Chapter 5 presents a new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO). It summarises a video-document pair into an extremely short summary consisting of one cover frame and one sentence. An unsupervised Hierarchical Optimal Transport Network (HOT-Net) is proposed that conducts learning across multiple modalities and uses optimal transport solvers to maximise semantic coverage. A new large-scale dataset is constructed to facilitate research in this new direction.

Chapter 6 presents a Topic-Guided Co-Attention Transformer method (TopicCAT) for XMSMO. It constructs a two-stage unsupervised learning strategy for unimodal and cross-modal modelling with clustering-based cross-modal topic guidance.

Chapter 7 concludes this thesis with a discussion of future work.

CHAPTER 2

Literature Review

The study investigates the task of text summarisation. In this chapter existing methods are reviewed in Section 2.1. Since most existing works on text summarisation are extractive, abstractive, and compressive-based, these approaches are discussed in Sections 2.1.1, 2.1.2, and 2.1.3 correspondingly.

Moreover, as the study evolves and investigates the task of extreme multimodal summarisation with multimodal summarisation, this chapter reviews multimodal summarisation in Section 2.3, with multimodal summarisation with unimodel output in 2.3.1 and that with multimodal output in Section 2.3.2. Existing extreme summarisation methods and video summarisation methods are also reviewed in Sections 2.4 and 2.2 respectively, since they are closely related to our study of extreme multimodal summarisation.

2.1 Text Summarisation

Text summarisation aims to condense a given document into a short and succinct summary that best covers the semantics of the document with the least redundancy. Most existing works on text summarisation are extractive, abstractive, and compressive-based. Firstly, Section 2.1.1, discusses existing extractive methods, which select salient sentences from a document to form its summary and ensure the production of grammatically and factually correct summaries. Secondly, Section 2.1.2, discusses existing abstractive methods, which usually formulate the task as a sequence-to-sequence generation task, with the document as the input sequence and the summary as the output sequence. Thirdly, in Section 2.1.3, existing compressive methods

are discussed. It is a recent approach which aims to select words, instead of sentences, from an input document to form a summary.

2.1.1 Extractive Text Summarisation

The extractive approach identifies and directly copies from the source document to form the summary. It has the advantage of ensuring grammatically correct and semantically meaningful summary output. However, it has the following drawbacks. Often the important information spread across multiple sentences; the extractive approach fails to capture unless extracting all these sentences. Also, without capturing the logical linkages between sentences, the ideas in the extracted summary may not be coherent [32]. Since it is relatively more straightforward, the majority of the previous works focused on extractive approach. In this section, we review existing extractive summarisation methods in two categories: non-learning based and learning-based methods.

2.1.1.1 Non-learning based Methods

Most of the non-learning based methods conceptualise text summarisation as a sentence ranking task. Each sentence in a given document is scored in terms of various sentence importance criteria, which measure how well the sentence could represent the document. The top-ranked sentences are combined to form a summary. These methods often heavily rely on handcrafted features in regards to linguistic knowledge by focusing on local and/or global contexts.

Local Context based Methods. Local context-based methods rank a sentence based on the features obtained from the sentence itself. Sentence features such as frequency-based and topic-based were studied. Frequency-based features [16, 35] assume that the occurrence of high-frequency terms in a sentence is associated with their importance. Topic-based features [46, 85, 55] assume that the density of a set of topic terms is highly correlated to the topic of a document.

Global Context based Methods. As local context features could overlook the correlations between sentences and lead to redundant summaries involving similar sentences, global context-based methods rank individual sentences from the perspective of the entire document. Discourse-based methods [71] construct a document's rhetorical structure and extract the sentences on the longest chain of the semantic structure, i.e. the main topic. Centroid-based methods [100] cluster the sentences of a document through similarity measures and rank the sentences based on their distances to the cluster centroids. TextRank [77], as a graph-based method, is the state-of-the-art non-learning based method. A graph among document sentences is first formed by connecting sentences using sentence similarity scores, then the sentence connectivity can be used to score the importance of a sentence. Nonetheless, the nature of these sentence based scoring methods could miss summary-level or document-level patterns.

2.1.1.2 Learning-based Methods

With the success of deep learning methods in the field of natural language processing (NLP), the deep learning-based text summarisation approach has then been popular in recent years. Similar to the non-learning based approach, the learning-based methods often conceptualise the task as a sentence ranking or a classification problem. For a sequence of sentences from the input document, each sentence is assigned with an importance score. Highly-ranked sentences are then combined to form a summary.

Most of these methods follow the sentence ranking conceptualisation, and a supervised encoder-decoder scheme is generally adopted [79, 134, 82, 123]. An encoder formulates document or sentence representations, and a decoder predicts a sequence of sentence importance scores with the supervision of ground-truth sentence labels. Since the ground-truth summaries are usually abstractive based and do not contain the labels of which sentences should be extracted, the training objective of these methods relies on creating proxy target labels for each sentence based on the similarity to the ground-truth summary sentence.

Instead of relying on proxy target labels, some supervised methods utilise reinforcement learning [81, 15, 66] by directly optimising the ROUGE metric, which is used as the training

reward. The reinforcement learning based summarisation task can be treated as a sentence ranking problem [81] similar to the aforementioned methods [79, 134, 82, 123] or as a contextual-bandit problem [66] . Instead of evaluating the scores of individual sentences, a contextual-bandit agent evaluates the reward by selecting a subset of sentences from a given document. The chosen sentences are then combined to form a summary.

To further address the discrepancy between the training objective and the evaluation metric, various unsupervised methods [138, 87] have been proposed to leverage pre-trained language models to compute sentence similarities and select important sentences. Some methods [138] use these similarities to construct a sentence graph and select sentences based on their centrality. Some methods [87] use these to score relevance and redundancy of sentences as selection criteria.

Although these learning-based methods have significantly improved summarisation performance in terms of the ROUGE metric, computationally expensive training costs are inevitable, and it is challenging to generalise the trained models to documents from other domains that have distributions different from the training dataset. Furthermore, the extraction result from both the existing learning-based and non-learning based approaches are lack of interpretability. It is difficult to explain the correspondence and the coverage between a summary and a source document using these deep models. Therefore, to address these limitations, it is necessary to revisit the non-learning based approach.

2.1.2 Abstractive Text Summarisation

The abstractive approach involves paraphrasing and natural language generation, which may generate novel words that are not in the source document. It is a better approximation of how human summarise texts, yet it is challenging to generate grammatically correct sentences. In this section, we review existing abstractive summarisation methods in two categories: non-learning based and learning-based methods.

2.1.2.1 Non-learning based Methods

The abstractive approach performs topic identification and interpretation by identifying and fusing important words of the input document. It performs summary generation in natural language by trying to ensure fluency and grammatical correctness of output summary. They are broadly categorised into template-based and graph-based.

Template-based Methods. Template-based methods assume that summary generation could be based on a set of pre-defined templates. At topic identification and interpretation stage, snippets are extracted from the input document. The summary is then generated by populating the snippets into the template.

Common methods to extract snippets from the input document include leveraging information extraction techniques [33] which extract snippets by linguistic extraction patterns and discourse-based approach [24] which aggregates the discourse trees of the sentences and extracts snippets from the relation tuples, and word graph [86].

The pre-defined template is usually based on the assumption of domain and topic covered by the method. Some methods experimented using a single template, assuming the output summaries follow universal structure [33] [24]. To have better generalisation to other domains, some [86] proposed to have multiple templates, such that output summaries could use the corresponding domain or topic-specific template based on the topic of the input document. Template-based methods suffer from poor generalisation since the predefined templates are domain-specific.

Graph-based Methods. To better generalise to different domains, graph-based methods assume that summary generation could be based on a traversal path on the graph-like structure constructed according to the input document, which does not limit to a specific domain. At topic identification and interpretation stage, common information among sentences is fused in the construction process to produce a graph structure, such as word tree, word graph and semantic graph. A set of predefined rules is then used to traverse the constructed graph. The traversed words are combined to form a summary in natural language.

The graph structure construction experimented include dependency trees [126] [108] and word graph [23] [20], with words represented as a node and adjacency relation as an edge. Semantic graphs are experimented to improve the semantic representation, such as rich semantic graph [95] [78], where the nodes represent nouns and verbs of the text and the edges represent the semantic relation between them, and abstract meaning representation (AMR) graph [57] [53], where the nodes represent concepts and the edges represent the relationship between them.

Traversal rules experimented include finding the shortest path between the 'start of sentence' node and the 'end of sentence' node [20], and searching the path with the total weight of word nodes [108]. For semantic graph, external AMR-to-text generator is experimented to convert an AMR graph into summary in natural language [53].

2.1.2.2 Learning-based methods

Learning-based abstractive methods formulate text summarisation as a sequence-to-sequence generation task, with the source document as the input sequence and the summary as the output sequence. Most existing methods follow the supervised RNN-based encoder-decoder framework [106, 131, 119, 61]. As supervised learning with ground-truth summaries may not provide useful insights on human judgment approximation, reinforcement training was proposed to optimise the ROUGE metric [92, 89], and to fine-tune a pre-trained language model [48]. These models naturally learn to integrate knowledge from the training data while generating an abstractive summary. Prior studies showed that these generative models are highly prone to external hallucination, thus may generate contents that are unfaithful to the original document [72].

2.1.3 Compressive Text Summarisation

Compressive methods select words from a given document to assemble a summary. Due to the lack of training dataset, not until recently there have emerged works for compressive summarisation [135, 75, 122, 13]. The formulation of compressive document summarisation is usually a two-stage extract-then-compress approach: it first extracts salient sentences from

a document, then compresses the extracted sentences to form its summary. Most of these methods are supervised, which require a parallel dataset with document-summary pairs to train. However, the ground-truth summaries of existing datasets are usually abstractive-based and do not contain supervision information needed for extractive summarisation or compressive summarisation. Several reinforcement learning based methods [135] use existing abstractive-based datasets for training, which is not aligned for compression. Note that existing compressors often perform compression sentence by sentence. As a result, the duplicated information among multiple sentences could be overlooked. Therefore, to address these limitations, we propose a novel unsupervised compressive method by exploring the dual-agent reinforcement learning strategy to mimic human judgment and perform text compression instead of sentence compression.

2.2 Video Summarisation

Video summarisation aims to summarise a video into keyframes that provide a compact yet informative representation of a video. Early researches [65, 140, 31] usually rely on handcrafted audiovisual features, such as audio, colour and motion, and utilise rule-based or clustering-based techniques to pick up keyframes . Recent researches [127, 68, 39, 67, 129] usually formulate this task as a sequence-to-sequence problem, with the video frames as the input sequence and the importance scores of each frame as the output sequence. The frames with the highest importance score are then selected as the keyframes. Most existing methods focus on modelling the temporal dependency and the spatio structure among frames [4]. The temporal dependency-based approach usually follows a supervised LSTM-based encoder-decoder framework [137, 118, 40] to exploit the temporal structure among the video frames and predict the importance of each frame. To further incorporate the knowledge of the spatio structure of frames, the spatiotemporal-based approach often adopts a combination of CNNs and RNNs [41, 37, 128], where pre-trained CNNs are used to represent the visual feature of the frames and RNNs are used to model the temporal dependency of the frames.

2.3 Multimodal Summarisation

Multimodal summarisation aims to condense an input with multiple modalities into a short summary [38] and creates concise and informative summaries that leverage the strengths of different modalities. Existing methods are commonly categorised by the modality of the output: with unimodal output and with multimodal output.

2.3.1 Multimodal Summarisation with Unimodal Output

Multimodal summarisation with unimodal output method summarises textual and visual inputs into a textual summary or keyframes. Existing research [9, 51, 88, 59] often focuses on generating better text summaries with the help of multimodal input. They often follow an encoder-decoder architecture, in which the encoder utilises pre-trained CNN networks for embedding visual features of images or video frames and RNN-based networks for encoding language encoding temporal dependency of video frames; the decoder utilises RNN-based networks for a text summary generation. A multimodal attention layer is used to fuse the multimodal representation.

2.3.2 Multimodal Summarisation with Multimodal Output

Multimodal summarisation with multimodal output methods usually summarise textual and visual inputs into a textual-visual summary. [141] first studied this task, which took a document and an image set as the input. A supervised attention-based encoder-decoder framework was devised. For encoding, a textual encoder and a visual encoder formulate the document and visual representations, respectively. For decoding, a textual decoder generates a textual summary, and a visual decoder selects the most representative image as a visual summary. Additionally, a multimodal attention layer was incorporated to fuse the textual and visual context information. To alleviate the modality-bias issue, a multitask learning was applied to jointly consider the two MSMO subtasks: summary generation and text-image relation recognition [142]. A hierarchical intra- and inter-modality correlation between the image and text inputs was studied to enhance the multimodal context representation [132].

[52] extended visual inputs to short videos, and introduced self-attentions to improve the multimodal context representation. Most existing MSMO methods are designed for short visual inputs, such as short videos and multiple images, without considering the summary length. Given the increasing pace of producing multimedia data and the subsequent challenge in keeping up with the explosive growth of such rich content, these existing methods may be sub-optimal to address the imminent issue of information overload of multimedia data.

2.4 Extreme Summarisation

Extreme summarisation is a form of summarisation which involves creating extremely concise summaries to further address the issue of information overload. Different from generic summarisation, extreme summarisation aims to generate extremely short summaries, such as cover images and one-line textual summaries, that users can browse and understand them at a glance. Existing extreme summarisation methods focus on unimodal input and output. They are reviewed in two categories, text-based and video-based.

2.4.1 Extreme Text Summarisation

The extreme text summarisation task was first explored by [80] who formulated the task as a sequence-to-sequence learning problem, where the input was a source document and the output was an extreme summary. A supervised encoder-decoder framework was studied and a topic model was incorporated as an additional input to involve the document-level semantic information and guide the summary to be consistent with the document theme. [7] introduced multi-task learning and incorporated the title generation as a scaffold task to improve the learning ability regarding the salient information in the document. These methods relied on integrating the knowledge from pre-trained embedding models to generate abstractive summaries. As a result, these generative models are highly prone to external hallucination and it is possible to generate contents unfaithful to the original document, which was shown by [72].

2.4.2 Extreme Video Summarisation

Extreme video summarisation methods can be conceptualized as a frame ranking task, which scores the frames in a video as the output. A deep learning method based on a CNN-based autoencoder architecture was first proposed [30], in which the training is unsupervised and the goal is to minimise a reconstruction loss considering the representativeness and aesthetic quality of the selected frames. The performance of different CNNs was compared and the ResNet-50 CNN outperformed the other CNNs, as studied by [98]. The scoring was improved by [102] by incorporating additional CNNs to consider the quality of faces. It utilised a Siamese CNN architecture, which was optimized by a piece-wise ranking loss using pairs of frames. [2] proposed a generative adversarial network that introduced a reinforcement learning scheme by rewarding the representativeness and aesthetic quality. Note that most of these methods encode a video as a sequence of frames directly, whilst the hierarchical semantic structure of a video has not been adequately explored.

CHAPTER 3

OTExtSum: Extractive Text Summarisation with Optimal Transport

---

In this chapter, we propose a novel non-learning-based method by for the first time formulating text summarisation as an Optimal Transport (OT) problem, namely Optimal Transport Extractive Summariser (OTExtSum). Optimal sentence extraction is conceptualised as obtaining an optimal summary that minimises the transportation cost to a given document regarding their semantic distributions. Such a cost is defined by the Wasserstein distance and used to measure the summary's semantic coverage of the original document. Comprehensive experiments on four challenging and widely used datasets - MultiNews, PubMed, BillSum, and CNN/DM demonstrate that our proposed method outperforms the state-of-the-art non-learning-based methods and several recent learning-based methods in terms of the ROUGE metric.

3.1 Introduction

A common practice for text summarisation is extractive summarisation which aims to select the salient sentences of a given document to form its summary. Extractive summarisation ensures the production of grammatically and factually correct summaries, though the output summaries could be inflexible. Since abstractive summaries are highly prone to contain contents that are unfaithful and nonfactual to the original document [72], extractive summaries are more practical for real-world scenarios, especially for the domains requiring formal writing such as legal, science, and journalism documents.

Existing methods [125] often first score the importance of individual sentences of a given document and then combine the top-ranked ones to form a summary. However, the sentences with high importance scores may not well represent the document from a global perspective,

which results in a sub-optimal summary. Recently, learning-based methods, especially those based on supervised and unsupervised deep learning techniques [81, 138, 134, 82, 123, 139, 87] can significantly improve summarisation performance. However, training deep learning models is computationally expensive, and it can be difficult to apply those models learned from a particular domain to other domains with different distributions. Moreover, deep learning methods generally lack interpretability for the summarisation process.

Motivated by these issues, we propose a novel non-learning based extractive summarisation method, namely Optimal Transport Extractive Summariser (OTExtSum). As illustrated in Figure 3.1, we formulate extractive summarisation based on the optimal transport (OT) theory [94]. A candidate summary can be evaluated by an OT plan regarding the optimal cost to transport between the semantic distributions of the summary and its original document. Then a Wasserstein distance can be obtained with this optimal plan to measure the discrepancy between the two distributions. To this end, it can be expected that a summary of high quality minimizes this Wasserstein distance. Moreover, a common assumption in the formulations of the OT problem is that the source and target distributions are fixed. In OTExtSum problem formulation, we relax this assumption by adding an extraction vector $\mathbf{m}^*$ to indicate which document sentences would be extracted to form the summary's semantic distribution, thus making the target distribution variable.

The semantic distributions of a given document and its candidate summary can be formulated in line with the frequency of their tokens. Inspired by Word Mover's Distance [47], summarisation can be conceptualized as moving the "semantics" of a given document to its summary, and the ideal summary is obtained at the minimal transportation cost. This ensures the highest semantic coverage of the given document and the least redundancy in the summary without explicitly modelling conventional criteria such as relevance and redundancy. Thus, under the OT plan, the Wasserstein distance indicates the candidate summary's semantic coverage of the given document.

We design two optimisation strategies to approximate the extraction vector $\mathbf{m}^*$, namely beam search strategy [113], which iteratively evaluates the semantic coverage scores of a set of candidate summaries to obtain the optimal extraction, and binary integer programming

strategy, which approximates the optimal extraction given the constraints of the Wasserstein distance and extraction budget. As a non-learning based method, OTExtSum does not require any training and is applicable to different document domains. Furthermore, it provides explainable results in terms of the semantic coverage of the summary.

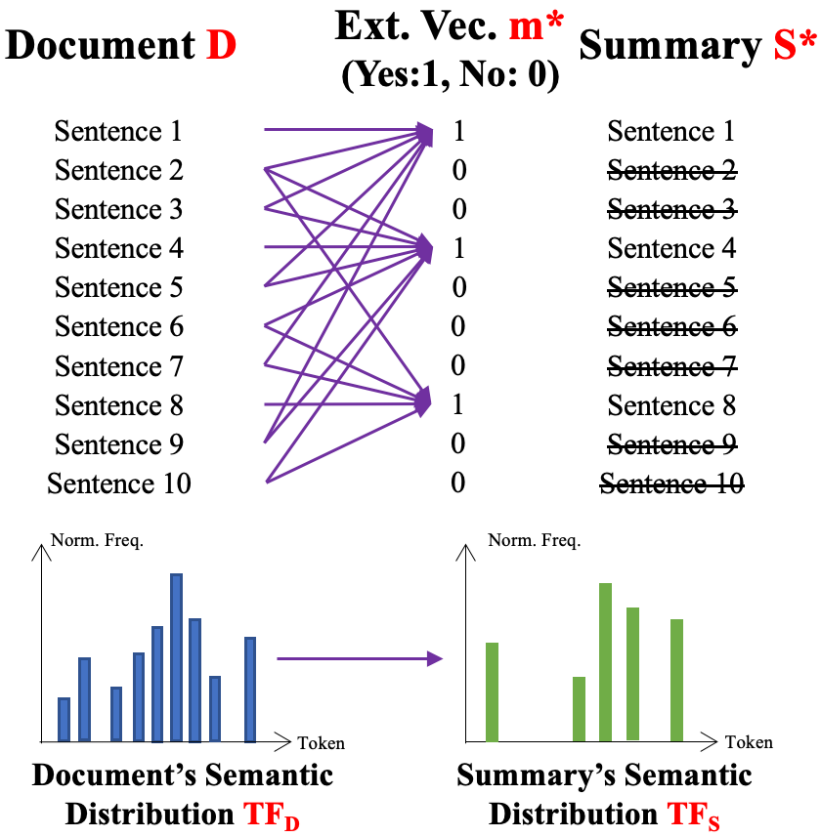

FIGURE 3.1. Illustration of Optimal Transport Extractive Summariser (OTExtSum).

Overall, the key contributions of this chapter are:

- We propose a non-learning based extractive summarisation method - OTExtSum by treating the text summarisation task as an optimal transport problem for the first time.
- We design two optimisation strategies for OTExtSum : beam search strategy and binary integer programming strategy.

- We present an interpretable visualisation of the semantic coverage of a generated summary by visualising the transport plan between summary tokens and document tokens.
- Comprehensive experimental results on four widely used datasets, including CNN/DM, MultiNews, BillSum and PubMed, demonstrate that OTExtSum outperforms the state-of-the-art non-learning based methods.

The remainder of this chapter is organised as follows. Section 3.2 describe the details of our proposed method. Section 3.3 presents comprehensive experiments to evaluate the effectiveness of our proposed method. Lastly, Section 3.4 concludes our study with discussions on our future work.

3.2 Proposed Method

As shown in Figure 3.1, OTExtSum utilizes a text OT approximation to obtain the optimal extraction vector $\mathbf{m}^* = [m_1, ..., m_n]^T$, where $m_i \in \{0, 1\}$ denotes whether the i-th sentence is to be extracted (denoted by 1) or not (denoted by 0). The optimal extraction vector $\mathbf{m}^*$ achieves an OT plan from the semantic distribution of the document to that of its optimal candidate summary which has the minimum total transportation cost.

The OT approximation consists of four components: 1) a tokeniser & embedding procedure that formulates token level representations and a semantic distribution estimation that computes the frequency of each token within a summary or a document ; 2) a transportation cost matrix that measures the cost using one token to represent another based on their Euclidean distances; 3) an OT solver that approximates Wasserstein distance and semantic coverage of the candidate summaries; and 4) an optimisation strategy that obtains the optimal extraction vector by choosing the summary with the minimum Wasserstein distance, and thus with the highest semantic coverage of the source document.

3.2.1 Optimal Transport

Consider a transportation problem that transports goods from a collection of suppliers $\mathbf{D} = \{d_i | i = 1, ..., N\}$ to a collection of customers $\mathbf{S} = \{s_j | j = 1, ..., N\}$, where d_i and s_j indicate the supply quantity of the i-th supplier and the order quantity of the j-th customer, respectively. Note that, in this study, we consider the number of suppliers to be the same as the customers. By defining t_{ij} as the quantity transported from the i-th supplier to the j-th customer, a transport plan $\mathbf{T} = \{t_{ij}\} \in \mathbf{R}^{N \times N}$ can be obtained. Given a cost matrix $\mathbf{C} = \{c_{ij}\} \in \mathbf{R}^{N \times N}$, where c_{ij} is the cost to deliver a unit of goods from the i-th supplier to the j-th supplier, the cost of the transport plan $\mathbf{T}$ can be calculated. Particularly, an OT plan $\mathbf{T}^* = \{t^*_{i,j}\} \in \mathbf{R}^{N \times N}$ in pursuit of minimising the transportation cost can be obtained by solving the following optimisation problem:

$$\begin{aligned} \mathbf{T}^* &= \underset{\mathbf{T}}{\text{argmin}} \sum_{i,j=1}^{N} t_{ij} c_{ij}, \\ \text{s.t.} \sum_{j=1}^{N} t_{ij} &= d_i, \quad \forall i \in \{1, ..., N\}, \\ \sum_{i=1}^{N} t_{ij} &= s_j, \quad \forall j \in \{1, ..., N\}, \\ t_{ij} &\geq 0, \quad \forall i, j \in \{1, ..., N\}, \end{aligned} \tag{3.1}$$

where the first two constraints indicate the quantity requirements for both suppliers and customers and the last constraint proves a non-negative order quantity. Mathematically, this OT problem is to find a joint distribution $\mathbf{T}$ with respect to a cost $\mathbf{C}$, of which the marginal distribution is $\mathbf{D}$ and $\mathbf{S}$. In particular, Wasserstein distance can be defined as:

$$d_W(\mathbf{D}, \mathbf{S} | \mathbf{C}) = \sum_{i,j} t^*_{i,j} c_{i,j}. \tag{3.2}$$

It can be viewed as the distance between the two probability distributions $\mathbf{D}$ and $\mathbf{S}$, if they are normalized, in line with the cost $\mathbf{C}$.

3.2.2 Semantic Distribution

In the context of text summarisation, denote $\mathbf{D} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ to represent a document, where $\mathbf{s}_i$ denotes the i-th sentence contained in the document. The sentence $\mathbf{s}_i$ has a semantic distribution $\text{TF}_i \in \mathbb{R}^N$ computed by the normalised bag-of-tokens with removal of stop-words:

$$\begin{aligned} \text{TF}_i &= [TF_{i1}, ..., TF_{iN}]^T, \\ TF_{ij} &= \frac{d_j}{\sum_{k=1}^{N} d_k}, \end{aligned} \tag{3.3}$$

where d_j indicates the count of the j-th token in a vocabulary of size N.

A document $\mathbf{D}$ has a semantic distribution $\text{TF}_{\mathbf{D}}$:

$$\text{TF}_{\mathbf{D}} = \frac{\text{TF}_1 + \ldots + \text{TF}_n}{n}. \tag{3.4}$$

For a summary $\mathbf{S} \subset \mathbf{D}$ with its corresponding extraction vector $\mathbf{m}$, of which the i-th element m_i is an indicator ($m_i = 1$ if $\mathbf{s}_i \in \mathbf{S}$, $m_i = 0$ otherwise), it has a semantic distribution $\text{TF}_{\mathbf{S}}$:

$$\text{TF}_{\mathbf{S}} = \frac{m_1 \times \text{TF}_1 + ... + m_n \times \text{TF}_n}{m_1 + ... + m_n}. \tag{3.5}$$

In our proposed method, a normalization step is introduced to approximate the semantic distributions of $\mathbf{D}$ and $\mathbf{S}$ with term frequency. Note that after the normalization, $\text{TF}_{\mathbf{D}}$ and $\text{TF}_{\mathbf{S}}$ have an equal total good quantities of 1 and can be *completely transported from one to the other*. In addition, $\text{TF}_{\mathbf{D}}$ and $\text{TF}_{\mathbf{S}}$ satisfy the property of discrete probability distributions, of which the sum should be 1.

3.2.3 Transport Cost between Tokens

We define the unit transportation cost between two tokens by measuring their semantic similarity. Intuitively, the more semantically dissimilar a pair of tokens are, the higher the "transport cost" of transporting one token to another. Given a pre-trained tokeniser and token

embedding model with N tokens, define $\mathbf{v}_i$ to represent the feature embedding of the i-th token. The transport cost from the i-th token to the j-th token c_{ij} in $\mathbf{C}$ can be written as:

$$c_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|_2 , \tag{3.6}$$

which is based on the Euclidean distance. [1]

3.2.4 Semantic Coverage of Candidate Summaries

Intuitively, a good summary $\mathbf{S}$ is supposed to be close to the document $\mathbf{D}$ in terms of their semantic distributions. OTExtSum utilizes the Wasserstein distance to measure the distance between the two associated semantic distributions $\text{TF}_{\mathbf{D}}$ and $\text{TF}_{\mathbf{S}}$ with the OT cost. The computation of the Wasserstein distance has time complexity of $O(p^3 \log(p))$ [1], where p denotes the number of unique words in the document.

In detail, it can be obtained with Eq. (3.2) as $d_W(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}}|\mathbf{C})$ with a pre-defined cost matrix $\mathbf{C}$. Then a semantic coverage score of the summary $\mathbf{S}$ in respect to the document $\mathbf{D}$ can be further defined based on the Wasserstein distance:

$$g(\mathbf{D}, \mathbf{S}) = 1 - d_W(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}}|\mathbf{C}). \tag{3.7}$$

Therefore, OTExtSum aims to search for an extraction vector $\mathbf{m}$, of which the corresponding summary $\mathbf{S}$ minimises the Wasserstein distance, i.e. maximising the semantic coverage score for the given document $\mathbf{D}$ by solving OT problems.

3.2.5 Optimisation Strategy

The remaining problem for OTExtSum is to search for the optimal extraction vector $\mathbf{m}^*$ which achieves the minimum total transportation cost from the semantic distribution of the document $\text{TF}_{\mathbf{D}}$ to that of the optimal summary $\text{TF}_{\mathbf{S}}$, given a budget B which is the number of sentences can be extracted to create a summary:

[1] We investigated the effect of different distance measurements. As discussed in Section 3.3.3, cost matrix based on the Euclidean distance and the cosine distance yield similar ROUGE scores.

$$\begin{aligned} \mathbf{m}^* = \underset{\mathbf{m}}{\text{argmin}}\ & d_W(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}}|\mathbf{C}), \\ & \text{s.t. } m_1 + ... + m_n \leq B. \end{aligned} \tag{3.8}$$

In search of optimal extraction vector $\mathbf{m}^*$, we design two optimisation strategies, namely beam search strategy to achieve better coverage approximation, and binary integer programming strategy to achieve better computational efficiency.

Algorithm 1 Optimisation of OTExtSum with Beam Search Strategy

Input : $\mathbf{D}$ the document, B the budget of the number of extracted sentences, K the beam width.

Output : $\mathbf{S}^*$ the optimal extractive summary.

Compute the cost matrix $\mathbf{C}$, and the document's semantic distribution $\text{TF}_{\mathbf{D}}$;
Initialise $\mathbf{m} = \mathbf{0}$, i.e. the candidate summary set $\mathbb{S} = \varnothing$;
while *# of sentences in candidate summary* $\leq B$ *;* **do** // Beam search
  for $k = 1, ..., |\mathbb{S}|$ **do**
    Generate the successor set $\mathbb{S}_b^k$ *for* $\mathbf{S}^k \in \mathbb{S}$*;*
  $\mathbb{S} \leftarrow \bigcup_k \mathbb{S}_b^k$;
  for $k = 1, ..., |\mathbb{S}|$ **do**
    Compute the semantic distribution $TF_{\mathbf{S}^k}$ *of* $\mathbf{S}^k \in \mathbb{S}$*;*
    Compute the Wasserstein distance $d_W(TF_{\mathbf{D}}, TF_{\mathbf{S}^k}|\mathbf{C})$ *and the semantic coverage* $g(TF_{\mathbf{D}}, TF_{\mathbf{S}^k}|\mathbf{C}))$*;*
  Keep the top K candidate summaries with the highest $g(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}^k}|\mathbf{C}))$ and prune the rest in $\mathbb{S}$;
$\mathbf{S}^* = \underset{\mathbf{S}^k \in \mathbb{S}}{\text{argmax}}\ g(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}^k}|\mathbf{C}))$;

3.2.5.1 Beam Search Strategy

The Beam Search (BS) strategy with the beam width K maintains the candidate summary set $\mathbb{S}$ and searches for the optimal extraction vector $\mathbf{m}^*$, thus the optimal extractive summary $\mathbf{S}^*$. Algorithm 1 presents the steps to obtain the optimal summary with OTExtSum using the BS strategy. The time complexity is $O(BKn(p^3\log(p)))$.

Initially, we have $\mathbf{m} = \mathbf{0}$, where none of the sentences are extracted. Then, each sentence in the document $\mathbf{D}$ is selected as a candidate summary, which derives a set of candidate extraction vectors corresponding to a set of candidate summaries, and its semantic coverage score can be evaluated. The top K candidate summaries in terms of the semantic coverage

are kept in the set $\mathbb{S}$ and the rest are pruned. During the b-th iteration of the beam search, by appending each possible sentence to an existing candidate summary $\mathbf{S}^k \in \mathbb{S}$, where the sentence is not in $\mathbf{S}^k$, a set of new candidate summaries $\mathbb{S}_b^k$ can be obtained. Then $\mathbb{S}$ is updated by combining all these sets of new candidate summaries in regards to k:

$$\mathbb{S} \leftarrow \bigcup_k \mathbb{S}_b^k. \tag{3.9}$$

At the end of beam search, a set of final K summary candidates within the budget B is obtained.

Among the K final candidates from the beam search, OTExtSum obtains the optimal extraction vector and thus the optimal summary by choosing the candidate with the highest semantic coverage of the document $\mathbf{D}$.

Algorithm 2 Optimisation of OTExtSum with Binary Integer Programming Strategy

Input : $\mathbf{D}$ the document, B the budget of the number of extracted sentences, T the number of iterations.
Output : $\mathbf{S}^*$ the optimal extractive summary.

Compute the cost matrix $\mathbf{C}$, Compute document's semantic distribution $\text{TF}_\mathbf{D}$;
Initialise $\mathbf{w} \in \mathbb{R}^n$;
for *iteration* $t \in [1, ..., T]$ **do**
  Convert $\mathbf{w}$ *to probability value* $\mathbf{pr}$ *with Sigmoid function;*
  Convert $\mathbf{pr}$ *to* $\mathbf{b} = [b_i, .., b_n]$ *by hard sampling from the Gumbel-Softmax distribution;*
  Construct summary's semantic distribution $TF_\mathbf{S}$*;*
  Compute the Wasserstein distance $d_W(TF_\mathbf{D}, TF_\mathbf{S}|\mathbf{C})$*;*
  Compute the L_1 *regularisation of* $\mathbf{b}$*;*
  Compute loss by weighted sum of the Wasserstein distance and the squared difference of B *and* $\mathbf{b}$*;*
  Compute gradients and update $\mathbf{w}$*;*
Compute $\mathbf{m}^*$ by soft sampling Sigmoid$(\mathbf{w})$ from the Gumbel-Softmax distribution;
Obtain $\mathbf{S}^*$ by extracting top-B sentences with the highest m_i values for $i = 1, ..., n$;

3.2.5.2 Binary Integer Programming Strategy

Some prior works showed that integer linear programming is an efficient solution to summarisation problem [73, 26].The Binary Integer Programming (BIP) strategy therefore is utilised to search for the optimal extraction vector $\mathbf{m}^*$ with T iterations. Based on the extraction vector, we obtain the optimal extractive summary $\mathbf{S}^*$. Algorithm 2 presents the optimisation

steps to obtain the optimal summary with OTExtSum using the BIP strategy. The time complexity is $O(T(p^3\log(p)))$.

As $\mathbf{m}^*$ is a multi-hot vector and is not differentiable, to make the backpropagation work, we optimise a proxy continuous vector $\mathbf{w} \in \mathbb{R}^n$, which is differentiable. Then we hard sample from the Gumbel-Softmax distribution [69] to discretise and compute a multi-hot vector $\mathbf{b}$ during the iterations, and soft sample to compute $\mathbf{m}^*$ at the end.

The BIP strategy optimises the following loss function w.r.t. $\mathbf{w}$, which is a weighted sum of the Wasserstein distance $d_W(\text{TF}_\mathbf{D}, \text{TF}_\mathbf{S})$ and the L_1 regularisation of $\mathbf{b}$ [2]:

$$d_W(\text{TF}_\mathbf{D}, \text{TF}_\mathbf{S}|\mathbf{C}) + \alpha|B - \sum_{i=1}^{n} b_i|, \tag{3.10}$$

where α denotes the weight of L_1 regularisation.

3.3 Experimental Results

3.3.1 Datasets

To validate the effectiveness of the proposed OTExtSum on the documents with various writing styles and its ability to achieve improved summarisation performance, we perform experiments on four widely used challenging datasets collected from different domains.

Dataset	**Multi-News**	**BillSum**	**PubMed**	**CNN/DM**
Domain	News	Law	Science	News
#Sent./Doc.	80	46	102	33
B	9	7	6	3
Test Set Size	5,622	3,269	6,658	11,490

TABLE 3.1. Overview of the datasets. #Sent./Doc. denotes the average number of sentences in the documents, B denotes the budget of number of extracted sentences.

CNN/DailyMail (CNN/DM) [34] is the standard single-document datasets with manually-written summaries. *Multi-News* [18] is a multi-document dataset which summarises multiple

[2]We choose L_1 regularisation for sparsity [84].

Method	**Multi-News**		
	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	42.3	14.2	22.4
ORACLE	45.4	20.6	28.1
Non-learning based Methods			
LexRank [17]	38.3	12.7	13.2
TextRank [77]	38.4	13.1	13.5
OTExtSum-BIP (GPT2)	40.6	12.1	20.7
OTExtSum-BIP (BERT)	40.6	12.1	20.7
OTExtSum-BS (Word2Vec)	42.3	12.8	21.9
OTExtSum-BS (GPT2)	42.4	**14.2**	**23.2**
OTExtSum-BS (BERT)	**43.1**	13.9	22.5
Unsupervised Deep Learning based Methods			
PacSum [138]	43.2	14.3	28.5
PMI [87]	40.5	13.2	19.8
Supervised Deep Learning based Method			
MatchSum [139]	46.2	16.5	41.9
PEGASUS [131]	47.5	18.7	24.9

TABLE 3.2. Comparisons between our OTExtSum and the state-of-the-art methods across different categories. The highest scores are **bold**, and the second highest ones are underlined.

Method	**BillSum**		
	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	43.5	25.6	37.8
ORACLE	43.7	25.7	38.0
Non-learning based Methods			
LSA [27]	32.6	15.7	26.3
TextRank [77]	34.4	17.8	27.8
OTExtSum-BIP (GPT2)	36.6	15.6	30.6
OTExtSum-BIP (BERT)	36.6	15.6	30.6
OTExtSum-BS (Word2Vec)	**40.1**	19.4	**34.3**
OTExtSum-BS (GPT2)	36.5	**19.7**	32.0
OTExtSum-BS (BERT)	37.5	**19.7**	32.6
Supervised Deep Learning based Method			
PEGASUS [131]	57.3	40.2	45.8

TABLE 3.3. Comparisons between our OTExtSum and the state-of-the-art methods across different categories. The highest scores are **bold**, and the second highest ones are underlined.

news articles. We concatenate the multiple articles as a single input. *BillSum* [44] is a dataset for law document summarization, which contains long state bill documents. *PubMed* [12] is a

Method	PubMed		
	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	34.0	8.6	27.1
ORACLE	37.1	15.5	30.4
Non-learning based Methods			
LSA [27]	33.9	9.9	29.7
LexRank [17]	39.2	**13.9**	**34.6**
OTExtSum-BIP (GPT2)	35.4	10.8	28.8
OTExtSum-BIP (BERT)	35.4	10.8	28.8
OTExtSum-BS (Word2Vec)	38.2	11.7	30.8
OTExtSum-BS (GPT2)	<u>39.7</u>	<u>13.8</u>	32.3
OTExtSum-BS (BERT)	**39.8**	13.6	32.3
Unsupervised Deep Learning based Methods			
PMI [87]	37.8	13.4	29.9
Supervised Deep Learning based Method			
MatchSum [139]	41.2	14.9	36.8
PEGASUS [131]	45.1	19.6	27.4

TABLE 3.4. Comparisons between our OTExtSum and the state-of-the-art methods across different categories. The highest scores are **bold**, and the second highest ones are <u>underlined</u>.

scientific article dataset that uses the abstract section as the ground-truth summary and the long body section as the document. Table 3.1 shows an overview of the four datasets.

We followed [139] to set B for CNN/DM, PubMed and Multi-News, and used the average number of sentences in the summaries to set B for BillSum since this is a common practice in the literatures [81]. These datasets were obtained from a source, namely HuggingFace Datasets [3]. Since OTExtSum does not require training, for a fair comparison, all experimental results are reported on the test splits of the four datasets only.

While *CNN/DM* contains shorter documents and summaries, the other three datasets are more challenging because they have more extended documents and summaries, thus having a higher chance to extract sentences containing redundant contents or having limited relevance to the document.

[3] https://huggingface.co/docs/datasets/

Method	CNN/DM		
	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	40.0	17.5	32.9
ORACLE	43.1	23.7	37.5
Non-learning based Methods			
TextRank [77]	34.1	**12.8**	22.5
OTExtSum-BIP (GPT2)	34.1	12.6	**28.1**
OTExtSum-BIP (BERT)	34.1	12.6	**28.1**
OTExtSum-BS (Word2Vec)	32.3	10.8	25.9
OTExtSum-BS (GPT2)	33.5	12.0	26.7
OTExtSum-BS (BERT)	**34.5**	**12.8**	27.8
Unsupervised Deep Learning based Methods			
PacSum [138]	40.3	17.6	24.9
PMI [87]	36.7	14.5	23.3
Supervised Deep Learning based Method			
MatchSum [139]	44.2	20.6	40.4
PEGASUS [131]	44.2	21.5	41.1

TABLE 3.5. Comparisons between our OTExtSum and the state-of-the-art methods across different categories. The highest scores are **bold**, and the second highest ones are underlined.

3.3.2 Implementation Details

In terms of the pre-trained token embedding model, we compare the static embedding model Word2Vec and the contextual embedding models BERT and GPT2. For the hyperparameter settings of the BIP strategy, the number of iteration T was set to 200, α was set to 1, and it used the SGD optimiser [110] with learning rate 0.1. For the BS strategy, the beam width K was set to 5 [4].

We obtained the pre-trained Word2vec (Google News 300 dimension) from GENSIM [5], and the contextual embedding models BERT (base version) and GPT2 from HuggingFace [6]. To compute the Wasserstein distances, we adopted GENSIM, the POT [7] and GeomLoss [19] libraries. List of stop-words was from NLTK library [8]. Our experiments were run on a

[4] We chose the beam width in line with a common practice in the literature [74]
[5] https://radimrehurek.com/gensim/index.html
[6] https://huggingface.co
[7] https://pythonot.github.io
[8] https://www.nltk.org

GeForce GTX 1080 GPU card. We obtain our ROUGE scores by using the pyrouge package [9].

Our OTExtSum is compared against LEAD [106], ORACLE [79], the state-of-the-art non-learning based methods and the recent unsupervised learning-based methods. LEAD and ORACLE are standard baselines in the summarisation task. LEAD baseline extracts the first several sentences of a document as a summary. ORACLE baseline greedily extracts the sentences that maximise the ROUGE-L score based on the reference summary. We compare with the results of strong non-learning-based methods, including LSA [27], TextRank [77], and LexRank [17]. Their results on MultiNews, BillSum, PubMed, and CNN/DM are from [18], [44], [12], and [87] respectively. For an informative reference, we report recent unsupervised learning-based methods, including PacSum [138], which its released model was trained on the news domain, and PMI [87], which its released models were trained on the news and science domains. Their results on CNN/DM are from [87]. Their results on MultiNews, BillSum, and PubMed are evaluated on the datasets with the corresponding released models from the same domains. And we include the results of the state-of-the-art supervised learning-based methods with extractive approach MatchSum from [139], and those with abstractive approach PEGASUS from [131].

3.3.3 Quantitative Analysis

The commonly used ROUGE metric [54] is also adopted for our quantitative analysis. It evaluates the content consistency between the generated summary and the reference summary. In detail, ROUGE-n scores measure the number of overlapping n-grams between the generated summary and the reference summary. A ROUGE-L score considers the longest common subsequence between the generated summary and the reference summary.

Performance Overview. The experimental results of OTExtSum on the four datasets are listed in Table 3.2, 3.3, 3.4 , 3.5 in terms of ROUGE-1, ROUGE-2 and ROUGE-L F-scores. We observed that the BS strategy could generally achieve better optimisation results than the

[9] https://pypi.org/project/pyrouge/

BIP strategy. It is in line with our design understanding that beam search can better reach the global optimum. Whereas, the two strategies achieve similar results in CNN/DM, which could be because CNN/DM has fewer document sentences and lower budget, thus fewer possible solutions and easier to find the optimum.

OTExtSum outperforms the state-of-the-art non-learning based methods and is comparable to the learning-based methods. Note that the state-of-the-art methods usually optimise at the sentence level, whilst OTExtSum is based on the summary level OT evaluation, by which the quality of the resulting summaries is improved.

We observed that OTExtSum obtains significantly better ROUGE scores than the baseline methods on Multi-News, BillSum and PubMed, while the improvement is not that significant on CNN/DM . When the summary is more extended, such as these three more challenging datasets, the summary sentences are more likely to have redundant content. That is, even summary-level optimisation is more difficult to achieve, our OTExtSum demonstrates higher improvements.

OTExtSum is a non-learning based method, and training is not required. Unlike learning-based methods, it is not limited by the training data domain and can be used for different domains. Experimental results demonstrate generalisation ability of OTExtSum over news, law, and science domains.

Method	**Multi-News**		
	ROUGE-1	ROUGE-2	ROUGE-L
Euc. \wo s.w.	43.1	13.9	22.5
Cos. \wo s.w.	43.1	13.9	22.5
Euc. \w s.w.	43.4	**14.4**	**23.4**
Cos. \w s.w.	**43.9**	14.2	23.1

TABLE 3.6. Ablation studies of OTExtSum based on the BS optimisation strategy and pre-trained BERT tokeniser.

Effects of Token Embeddings Models. OTExtSum is dependent on a pre-trained token embedding method. Specifically, the token embedding model affects the cost matrix C and the tokenisation, thus the frequency vector, of the document. We examine how different

Method	**BillSum**		
	ROUGE-1	ROUGE-2	ROUGE-L
Euc. \wo s.w.	37.5	**19.7**	32.6
Cos. \wo s.w.	**39.0**	19.5	**33.6**
Euc. \w s.w.	36.9	19.6	32.2
Cos. \w s.w.	38.1	19.6	33.0

TABLE 3.7. Ablation studies of OTExtSum based on the BS optimisation strategy and pre-trained BERT tokeniser.

Method	**PubMed**		
	ROUGE-1	ROUGE-2	ROUGE-L
Euc. \wo s.w.	39.8	13.6	32.2
Cos. \wo s.w.	39.8	13.6	32.3
Euc. \w s.w.	**40.6**	**13.8**	**33.0**
Cos. \w s.w.	**40.6**	13.6	32.9

TABLE 3.8. Ablation studies of OTExtSum based on the BS optimisation strategy and pre-trained BERT tokeniser.

Method	**CNN/DM**		
	ROUGE-1	ROUGE-2	ROUGE-L
Euc. \wo s.w.	**34.5**	**12.8**	**27.8**
Cos. \wo s.w.	34.4	12.4	27.7
Euc. \w s.w.	34.1	12.1	27.1
Cos. \w s.w.	34.1	12.1	27.1

TABLE 3.9. Ablation studies of OTExtSum based on the BS optimisation strategy and pre-trained BERT tokeniser.

token embedding models would affect the performance of OTExtSum by comparing static embedding model Word2Vec, and contextual embedding models BERT and GPT2.

The results on most of the datasets indicate that a more advanced contextual embedding model such as BERT and GPT2 is more effective than a static embedding model Word2Vec. It is in line with the intuitive understanding that a more representative model with adequate training samples often approximates better token embeddings and representation. Despite that, the performance of OTExtSum with Word2Vec is surprisingly competitive.

Effects on Stop-words. We investigate the impact of stop-words on the performance of OTExtSum. As shown in Table 3.6, 3.7, 3.8, 3.9 (s.w. denotes stop-words), the effect varies

slightly across the datasets, and may not much influence the ROUGE scores. It could be because text summarisation does not generally depend on stop-words. A side benefit of removing the stop-words is reducing the vocabulary size and thus the computation time of OT.

Effects on Distance Measurement. We examine how the distance measurement of the cost matrix would impact the performance of OTExtSum. As shown in Table 3.6, 3.7, 3.8, 3.9 (Euc. denotes the Euclidean distance and Cos. denotes the cosine distance), cost matrix based on the cosine distance and the Euclidean distance usually yield similar ROUGE scores.

3.3.4 Interpretable Visualisation

OTExtSum is able to provide an interpretable visualisation of the summarisation procedure. Figure 3.2 illustrates the transport plan heatmap, which indicates the transportation of semantic contents between tokens in the document and its resulting summary. The higher the intensity, the more the semantic content of a particular document token is covered by a summary token. Purple line highlights the transportation from the document to the summary of semantic content of token "month", which appears in both the document and the summary. Red line highlights how the semantic content of token "sponsor", which appears in the document only but not the summary, are transported to token "tour" and "extension", which are semantically closer and have lower transport cost, and thus achieve a minimum transportation cost in the OT plan.

3.3.5 Qualitative Analysis

Figure 3.3 , 3.4 , 3.5, and 3.6 compare the summaries produced by OTExtSum and TextRank. TextRank extracted sentences that are salient on their own yet redundant when combined to form a summary. In comparison, OTExtSum is able to compose summaries that have higher semantic coverage and less redundant content.

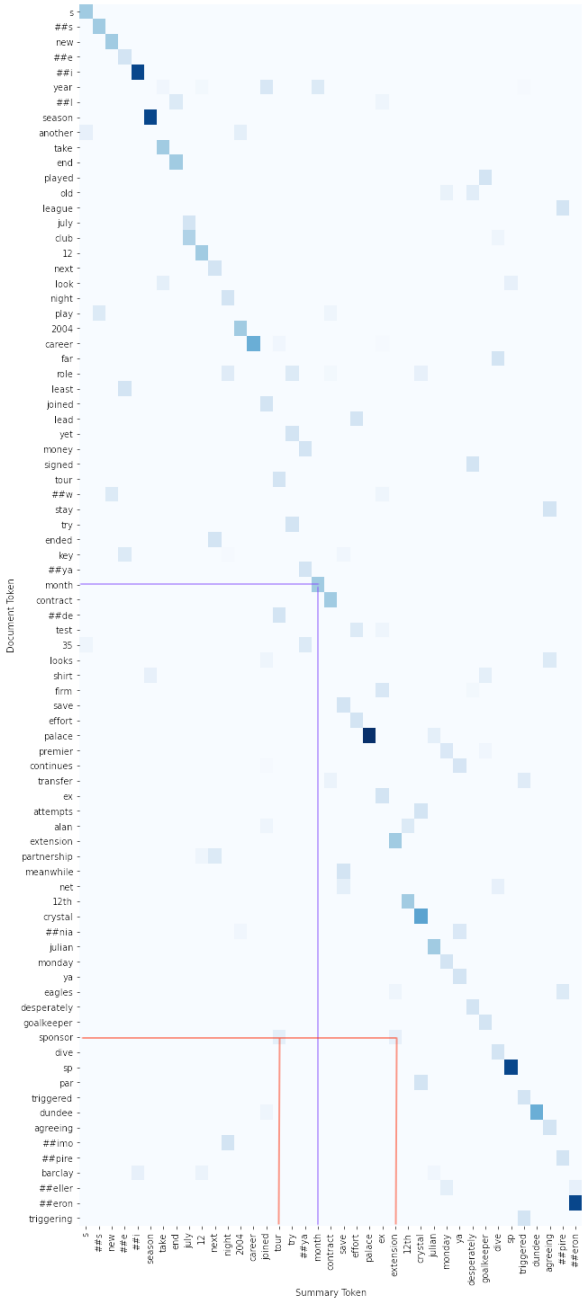

FIGURE 3.2. Interpretable visualisation of the OT plan from a source document to a resulting summary on the CNN/DM dataset.

Figure 3.3 shows a sample summary comparison on the **Multi-News** dataset. OTExtSum based summary sentences are highlighted in yellow colour. TextRank based summary sentences are underlined in red colour. TextRank extracted redundant contents, specifically the part (1) is duplicated with the part (3), and the part (2) is duplicated with the part (4). The

Source Document:

WASHINGTON – (1) **Sen. John McCain (R-Ariz.) said Wednesday night he will no longer support Richard Mourdock, unless the Indiana Senate candidate apologizes for his recent comments on rape.** (...) "If he apologizes and says he misspoke and he was wrong and he asks the people to forgive him, then obviously I'd be the first. (...)McCain's call for an apology comes as Mitt Romney faces mounting pressure from Democrats to withdraw his endorsement of Mourdock and have a TV ad he starred in for the Indiana state treasurer removed from the airwaves.(...) "I think it depends on what he does," McCain told CNN's Anderson Cooper, when asked if he still counts himself among Mourdock's supporters. (...) I think that even when life begins in that horrible situation of rape, that it is something that God intended to happen." (2) **Romney's campaign confirmed Wednesday that he continues to support Mourdock and has not asked that the ad be pulled.** Romney campaign spokeswoman Andrea Saul did seek to distance the GOP nominee from the views Mourdock expressed." Gov. Romney disagrees with Richard Mourdock, and Mr. Mourdock's comments do not reflect Gov. Romney's views," Saul said in an emailed statement. "We disagree on the policy regarding exceptions for rape and incest but still support him." Mourdock declined to apologize during a press conference on Wednesday, saying instead that he regretted his words had been misinterpreted. UPDATE: Oct. 25 – McCain spokesman Brian Rogers on Thursday clarified (3) **McCain's earlier comments that he would withdraw support for Mourdock if he did not apologize**, saying that the Arizona senator still supports the Indiana Senate candidate. "Senator McCain was traveling yesterday in Florida and did not have an opportunity to see Mr. Mourdock's full press conference before he taped his CNN interview," Rogers said, referring to the Wednesday conference where Mourdock said he regretted that he had been misinterpreted. "Senator McCain is glad that Mr. Mourdock apologized to the people of Indiana and clarified his previous statement. Senator McCain hopes the people of Indiana will elect Mr. Mourdock to the U.S. Senate." President Obama on Wednesday criticized Senate candidate Richard Mourdock's controversial comments on rape, saying that the Indiana Republican was wrong when he called pregnancy resulting from rape "something God intended to happen." "I don't know how these guys come up with these ideas," Obama said during a taping of "The Tonight Show" in Burbank, Calif. "Let me make a very simple proposition. Rape is rape. It is a crime." "These various distinctions about rape don't make too much sense to me, don't make any sense to me," he said. (Also on POLITICO: Exclusive interview: Mourdock braces for fallout) Obama also laid out a contrast between his position on women's issues and Mitt Romney's, saying that "Roe vs. Wade is probably hanging in the balance" depending on the outcome of the presidential race and the winner's eventual Supreme Court appointments. Obama added that Mourdock's remarks are evidence of why government shouldn't make decisions on women's health. "This is exactly why you don't want a bunch of politicians, mostly male, making decisions about women's health care decisions. Women are capable of making these decisions in consultation with their partners, with their doctors," he said to loud applause. "And for politicians to want to intrude in this stuff, oftentimes without any information, is a huge problem. And this is obviously part of what's stake at this election." Mourdock's comment, made Tuesday, has mobilized the Obama campaign and Democrats to call on Mitt Romney to distance himself from the Senate candidate. (Also on POLITICO: GOP splits over Mourdock comment) (4) **The Romney campaign distanced itself from Mourdock's statement on rape, but still supports the candidate and has not asked him to stop using an ad featuring Romney.** "Gov. Romney disagrees with Richard Mourdock, and Mr. Mourdock's comments do not reflect Gov. Romney's views. We disagree on the policy regarding exceptions for rape and incest but still support him," press secretary Andrea Saul said. (...) Sen. John McCain (R-Ariz.) said his support for the Indiana Republican "depends on what he does." "If he apologizes and says he misspoke and he was wrong and asks the people to forgive him, I would be the first" to support him, McCain said on CNN's "Anderson Cooper 360." Earlier Wednesday, the Obama campaign released a web video juxtaposing clips from Romney's pro-Mourdock ad with Mourdock's remarks on rape. The video ends with message, "Mitt Romney: extremely conservative to this day." Read more about: Richard Mourdock

Reference Summary:

President Obama slammed Senate candidate Richard Mourdock's rape comments on the Tonight Show last night. "I don't know how these guys come up with these ideas," Obama said. "Let me make a very simple proposition. Rape is rape. It is a crime." He noted "Roe vs. Wade is probably hanging in the balance" of the election, Politico reports. Mourdock's comments that pregnancies from rape are "something God intended" are "exactly why you don't want a bunch of politicians, mostly male, making decisions about women's health care decisions," Obama said. "Women are capable of making these decisions in consultation with their partners, with their doctors." Meanwhile, asked by Anderson Cooper whether he still supports Mourdock, John McCain said it "depends on what he does," the Huffington Post reports. "If he apologizes and says he misspoke and he was wrong and he asks the people to forgive him, then obviously I'd be the first" to support him. Mourdock has in fact issued an apology. Mitt Romney, for his part, still supports Mourdock, though his campaign says he "disagrees" with the Indiana Republican's comments: "We disagree on the policy regarding exceptions for rape and incest but still support him." The campaign says it hasn't called on Mourdock to withdraw ads featuring Romney.

FIGURE 3.3. A sample summary comparison on the Multi-News dataset.

summary generated by OTExtSum has ROUGE-1 F-Score of 65.21 and Semantic Coverage Score of 0.93, while the summary generated by TextRank has ROUGE-1 F-Score of 44.87 and Semantic Coverage Score of 0.89. Semantic Coverage Score of the ground-truth summary is 0.89.

Figure 3.4 shows a sample summary comparison on the **BillSum** dataset. OTExtSum based summary sentences are highlighted in yellow colour. TextRank based summary sentences are underlined in red colour. TextRank extracted redundant contents, specifically the part (1), (2) (3), (4), and (5) are duplicated. The summary generated by OTExtSum has ROUGE-1 F-Score of 44.2 and Semantic Coverage Score of 0.92, while the summary generated by

Source Document:

SECTION 1. SHORT TITLE; TABLE OF CONTENTS. (a) Short Title.–This Act may be cited as the "The Water Recycling and Riverside-Corona Feeder Act of 2006". (...) (1) **Inland Empire and Cucamonga Valley recycling projects.** TITLE II–PROJECTS IN RIVERSIDE AND SAN BERNARDINO COUNTIES Sec. 201. Planning, design, and construction of the Riverside-Corona Feeder. Sec. 202. Project authorizations. TITLE I–THE INLAND EMPIRE REGIONAL WATER RECYCLING INITIATIVE SEC. 102. SHORT TITLE. This title may be cited as the "The Inland Empire Regional Water Recycling Initiative". SEC. 103. INLAND EMPIRE AND CUCAMONGA VALLEY RECYCLING PROJECTS. (a) Recycling Projects.–The Reclamation Wastewater and Groundwater Study and Facilities Act (Public Law 102-575, Title XVI; 43 U.S.C. 390h et seq.) is amended by adding at the end the following: "SEC. 1637. (2) **INLAND EMPIRE REGIONAL WATER RECYCLING PROJECT.** "(...) (3) construction of **the Inland Empire regional water recycling project** (...) (d) Authorization of Appropriations.–There is authorized to be appropriated to carry out this section $20,000,000. "SEC. 1638. (4) **CUCAMONGA VALLEY WATER RECYCLING PROJECT.** (...) (5) **Cucamonga Valley Water Recycling Project.** (...) plan, design, and construct a water supply project, the Riverside-Corona Feeder, (...) the cost to plan, design, and construct the project described (...) (a) In General.–The Reclamation Wastewater and Groundwater Study and Facilities Act (Public Law 102-575, title XVI; 43 U.S.C. 390h et seq.) (...) (b) Cost Sharing.–The Federal share of the cost of the project described in subsection (a) shall not exceed 25 percent of the total cost of the project. (...)

Reference Summary:

Water Recycling and Riverside-Corona Feeder Act of 2006 - Inland Empire Regional Water Recycling Initiative - Amends the Reclamation Wastewater and Groundwater Study and Facilities Act to authorize the Secretary of the Interior: (1) in cooperation with the Inland Empire Utilities Agency, to participate in the design, planning, and construction (design) of the Inland Empire regional water recycling project, California; (2) in cooperation with the Cucamonga Valley Water District, to participate in the design of the Cucamonga Valley Water District satellite recycling plants in Rancho Cucamonga to reclaim and recycle approximately two million gallons per day of domestic wastewater; (3) in cooperation with the Yucaipa Valley Water District, to participate in the design of projects to treat impaired surface water, reclaim and reuse impaired groundwater, and provide brine disposal within the Santa Ana Watershed; and (4) in cooperation with the City of Corona Water Utility, to participate in the design of, and land acquisition for, a project to reclaim and reuse wastewater, including degraded groundwaters, within and outside of the City. Limits the federal cost share of each project to 25%. Authorizes the Secretary, in cooperation with the Western Municipal Water District, to participate in a project to design the Riverside-Corona Feeder, which includes 20 groundwater wells and 28 miles of pipeline in San Bernardino and Riverside Counties, California. Limits the federal share of the project design and planning study costs.

FIGURE 3.4. A sample summary comparison on the BillSum dataset.

TextRank has ROUGE-1 F-Score of 33.2 and Semantic Coverage Score of 0.77. Semantic Coverage Score of the ground-truth summary is 0.84.

Figure 3.5 shows a sample summary comparison on the **PubMed** dataset. OTExtSum based summary sentences are highlighted in yellow colour. TextRank based summary sentences are underlined in red colour. TextRank extracted redundant contents, specifically the part (1) is duplicated with the part (4), and the part (2) is duplicated with the part (3). The summary generated by OTExtSum has ROUGE-1 F-Score of 73.1 and Semantic Coverage Score of 0.92, while the summary generated by TextRank has ROUGE-1 F-Score of 66.0 and Semantic Coverage Score of 0.89. Semantic Coverage Score of the ground-truth summary is 0.91.

Figure 3.6 shows a sample summary comparison on the **CNN/DM** dataset. OTExtSum based summary sentences are highlighted in yellow colour. TextRank based summary sentences are underlined in red colour. TextRank extracted redundant contents, specifically the part (1) is duplicated with the part (2). The summary generated by OTExtSum has ROUGE-1 F-Score of 50.5 and Semantic Coverage Score of 0.89, while the summary generated by TextRank has ROUGE-1 F-Score of 35.7 and Semantic Coverage Score of 0.83. Semantic Coverage Score of the ground-truth summary is 0.80.

Source Document:

to report (1) **the effective treatment of radiation macular edema following ruthenium-106 plaque brachytherapy for a choroidal melanoma with a dexamethasone 0.7-mg (ozurdex) intravitreal implant** . a 65-year - old caucasian woman was suffering from radiation macular edema following ruthenium-106 plaque brachytherapy for a choroidal melanoma on her left eye . (...) (2) **seven months after the development of radiation macular edema , she received a single intravitreal injection of dexamethasone 0.7 mg (ozurdex) .** four weeks following the injection , her best - corrected visual acuity improved from 0.3 to 0.5 . (...) other studies suggest rates of radiation maculopathy from plaque radiotherapy of 18% , 23% , and 42.8% . (...) recently , a sustained - release dexamethasone implant (ozurdex) proved to be effective for the treatment of macular edema secondary to a variety of underlying diseases with a potentially lower rate of adverse events . there are no cases of radiation macular edema after ruthenium (ru)-106 plaque brachytherapy for choroidal melanoma resolved by an intravitreal dexamethasone 0.7-mg implant described so far in the literature . we report a case of radiation macular edema after ru-106 brachytherapy for a choroidal melanoma , refractory to a previous treatment with intravitreal bevacizumab , and resolved with significant improvement of visual function following an intravitreal injection of dexamethasone 0.7 mg . (...) she underwent one intravitreal injection of 0.5 mg bevacizumab (avastin , genentech / roche) in the following months without functional or anatomical improvement . (3) **seven months after the development of radiation macular edema , she received a single intravitreal injection of dexamethasone 0.7 mg (ozurdex)** as off - label treatment . (...) one of the main mechanisms of the chronic macular edema is the alteration of muller cells functionality ; it has been experimentally shown that steroids , by reducing the osmotic swelling of the muller 's cells , improve their functionality and reduce the macular edema . this could indicate that (4) **dexamethasone implant (ozurdex) might be an effective treatment option** not only in retinal vein occlusion and noninfectious uveitis , but can also be considered as off - label treatment **in radiation macular edema after ru-106 plaque brachytherapy for choroidal melanoma .**

Reference Summary:

purposeto report the effective treatment of radiation macular edema following ruthenium-106 plaque brachytherapy for a choroidal melanoma with a dexamethasone 0.7-mg (ozurdex) intravitreal implant.methodsan interventional case report with optical coherence tomography (oct) scans.resultsa 65-year - old caucasian woman was suffering from radiation macular edema following ruthenium-106 plaque brachytherapy for a choroidal melanoma on her left eye . she had undergone one intravitreal injection of 0.5 mg bevacizumab (avastin , genentech / roche) in the following months without functional or anatomical improvement . seven months after the development of radiation macular edema , she received a single intravitreal injection of dexamethasone 0.7 mg (ozurdex) . four weeks following the injection , her best - corrected visual acuity improved from 0.3 to 0.5 . radiation macular edema resolved with a reduction of central retinal thickness from 498 m before ozurdex injection to 224 m after ozurdex injection , as measured by oct scan.conclusiondexamethasone 0.7 mg (ozurdex) has proven to be an effective treatment option in retinal vein occlusion and noninfectious uveitis . it can also be considered as off - label treatment in radiation macular edema following ruthenium-106 plaque brachytherapy for a choroidal melanoma .

FIGURE 3.5. A sample summary comparison on the PubMed dataset.

Source Document:

(1) **Julian Speroni will take his Crystal Palace career into a 12th season** after agreeing a new contract. The goalkeeper, who joined from Dundee in July 2004, has triggered a 12-month extension that will expire at the end of next season. Speroni, yet again, has played a lead role for Palace this season in the club's attempts to stay in the Barclays Premier League. Julian Speroni will be at Crystal Palace for another year at least after triggering a 12-month contract extension . Speroni desperately dives to try and save Yaya Toure's effort against Crystal Palace on Monday night . The 35-year-old will have a testimonial against Dundee at the end of the season. But his Eagles career looks far from over as he continues to play a key role for Alan Pardew. Meanwhile, Palace are on the look out for a new shirt sponsor after money transfer firm Neteller ended their partnership with the club. Speroni signed from Dundee in 2004 and (2) **will take his Palace career into a 12th season .**

Reference Summary:

Julian Speroni signed from Dundee in 2004 and is a Crystal Palace legend . Argentine goalkeeper triggers contract extension at Selhurst Park . Speroni will have testimonial at the end of the season against Dundee .

FIGURE 3.6. A sample summary comparison on the CNN/DM dataset.

3.4 Conclusion

In this chapter, we have presented OTExtSum, the first optimal transport-based optimisation method for extractive text summarisation. It aims to identify an optimal subset of sentences for producing a summary that achieves high semantic coverage of the document by minimising the Wasserstein distance between the semantic distributions of the document and the summary. It helps obtain a summary from a global perspective and provides an interpretable visualisation of extraction results. In addition, OTExtSum does not require computationally expensive training. The comprehensive experiments demonstrate the effectiveness of OTExtSum, which

is generalisable over various document domains. In our future work, we will explore other OT solvers for extractive summarisation.

CHAPTER 4

Efficient and Interpretable Compressive Text Summarisation with Unsupervised Dual-Agent Reinforcement Learning

---

In this chapter, we propose an efficient and interpretable compressive summarisation method that utilises unsupervised dual-agent reinforcement learning to optimise a summary's semantic coverage and fluency by simulating human judgment on summarisation quality. Our model consists of an extractor agent and a compressor agent, and both agents have a multi-head attentional pointer-based structure. The extractor agent first chooses salient sentences from a document, and then the compressor agent compresses these extracted sentences by selecting salient words to form a summary without using reference summaries to compute the summary reward. To our best knowledge, this is the first work on unsupervised compressive summarisation. Experimental results on three widely used datasets (e.g., Newsroom, CNN/DM, and XSum) show that our model achieves promising performance and a significant improvement on Newsroom in terms of the ROUGE metric, as well as interpretability of semantic coverage of summarisation results.

4.1 Introduction

Compressive summarisation is a recent approach which aims to select words, instead of sentences, from an input document to form a summary, which improves the factuality and conciseness of a summary. The formulation of compressive document summarisation is usually a two-stage extract-then-compress approach [135, 75, 122, 13]: it first extracts salient sentences from a document, then compresses the extracted sentences to form its summary. Most of these methods are supervised, which require a parallel dataset with

document-summary pairs to train. However, the ground-truth summaries of existing datasets are usually abstractive-based and do not contain supervision information needed for extractive summarisation or compressive summarisation [122, 75, 13].

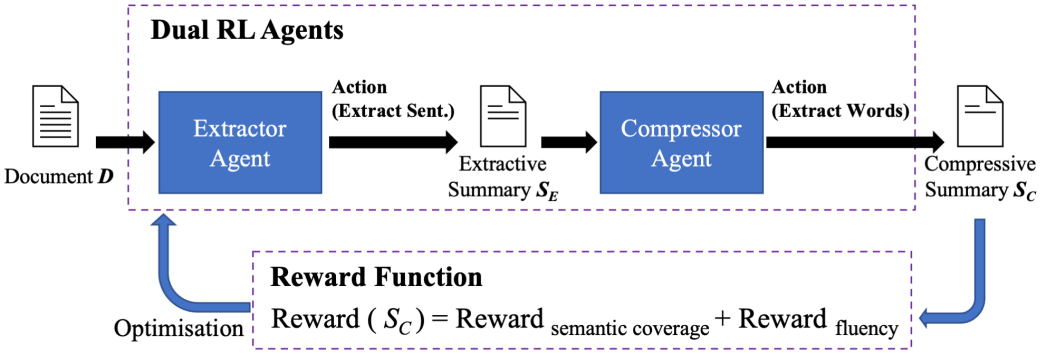

FIGURE 4.1. Illustration of our proposed URLComSum.

Therefore, to address these limitations, we propose a novel unsupervised compressive summarisation method with dual-agent reinforcement learning strategy to mimic human judgment, namely URLComSum. As illustrated in Figure 4.1, URLComSum consists of two modules, an extractor agent and a compressor agent. We model the sentence and word representations using an efficient Bi-LSTM [28] with multi-head attention [115] to capture both the long-range dependencies and the relationship between each word and each sentence. We use a pointer network [117] to find the optimal subset of sentences and words to be extracted since the Pointer Network is well-known for tackling combinatorial optimization problems. The extractor agent uses a hierarchical multi-head attentional Bi-LSTM model for learning the sentence representation of the input document and a pointer network for extracting the salient sentences of a document given a length budget. To further compress these extracted sentences all together, the compressor agent uses a multi-head attentional Bi-LSTM model for learning the word representation and a pointer network for selecting the words to assemble a summary.

As an unsupervised method, URLComSum does not require a parallel training dataset.We propose an unsupervised reinforcement learning training procedure to mimic human judgment: to reward the model that achieves high summary quality in terms of semantic coverage and language fluency. Inspired by Word Mover's Distance [47], the semantic coverage reward is measured by Wasserstein distance [94] between the semantic distribution of the document and

that of the summary. The fluency reward is measured by Syntactic Log-Odds Ratio (SLOR) [91]. SLOR is a referenceless fluency evaluation metric, which is effective in sentence compression [42] and has better correlation to human acceptability judgments [49].

The key contributions of this chapter are:

- We propose the first unsupervised compressive summarisation method with dual-agent reinforcement learning, namely URLComSum.
- We design an efficient and interpretable multi-head attentional pointer-based neural network for learning the representation and for extracting salient sentences and words.
- We propose to mimic human judgment by optimising summary quality in terms of the semantic coverage reward, measured by Wasserstein distance, and the fluency reward, measured by Syntactic Log-Odds Ratio (SLOR).
- Comprehensive experimental results on three widely used datasets, including CNN / DM, XSum, Newsroom, demonstrate that URLComSum achieves great performance.

The remainder of this chapter is organised as follows. Section 4.2 describe the details of our proposed method. Section 4.3 presents comprehensive experiments to evaluate the effectiveness of our proposed method. Lastly, Section 4.4 concludes our study with discussions on our future work.

4.2 Proposed Method

As shown in Figure 4.1, our proposed compressive summarisation method, namely URLComSum, consists of two components, an extractor agent and a compressor agent. Specifically, the extractor agent selects salient sentences from a document $\mathbf{D}$ to form an extractive summary $\mathbf{S_E}$, and then the compressor agent compresses $\mathbf{S_E}$ by selecting words to assemble a compressive summary $\mathbf{S_C}$.

4.2.1 Extractor Agent

Given a document $\mathbf{D}$ consisting of a sequence of M sentences $\{\mathbf{s}_i|i = 1, ..., M\}$, and each sentence $\mathbf{s}_i$ consisting of a sequence of N words $\{\mathbf{we}_{ij}|j = 1, ..., N\}$[1], the extractor agent aims to produce an extractive summary $\mathbf{S_E}$ by learning sentence representation and selecting L_E sentences from $\mathbf{D}$. As illustrated in Figure 4.2, we design a hierarchical multi-head attentional sequential model for learning the sentence representations of the document and using a Pointer Network to extract sentences based on their representations.

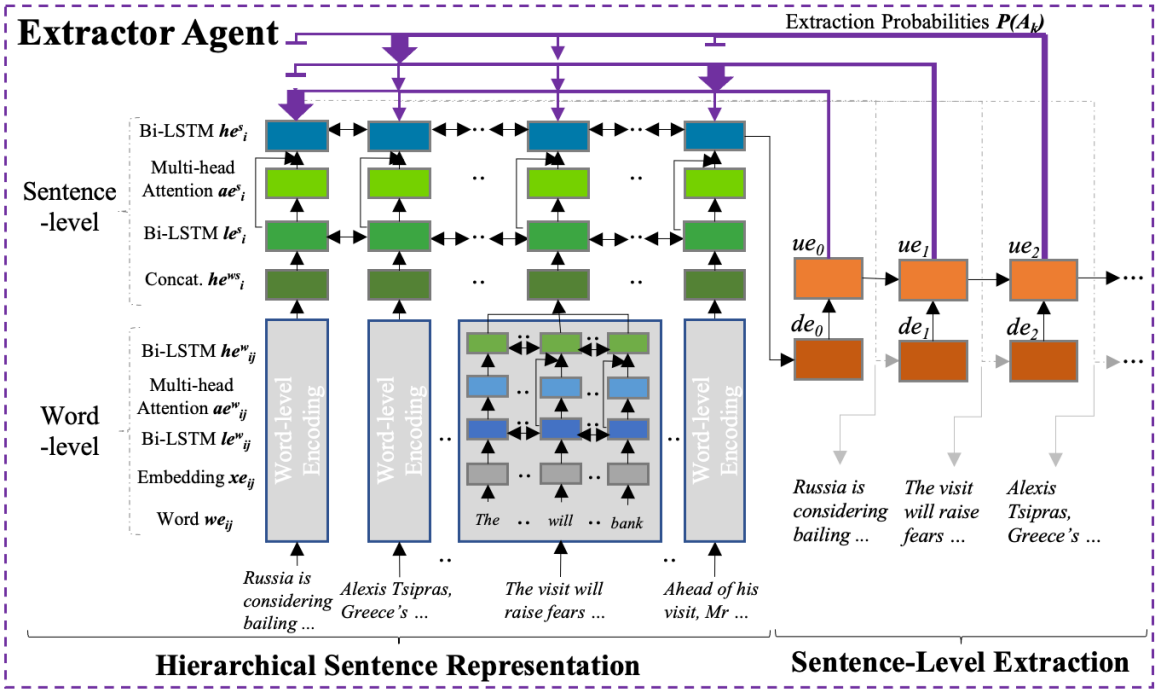

FIGURE 4.2. Illustration of the extractor agent.

4.2.1.1 Hierarchical Sentence Representation

To model the local context of each sentence and the global context between sentences, we use two-levels Bi-LSTMs to model this hierarchical structure, one at the word level to encode the word sequence of each sentence, one at the sentence level to encode the sentence sequence of the document. To model the context-dependency of the importance of words and sentences,

[1]We have pre-fixed the length of each sentence and each document by padding.

we apply two levels of multi-head attention mechanism [115], one at each of the two-level Bi-LSTMs.

Given a sentence $\mathbf{s}_i$, we encode its words into word embeddings $\mathbf{xe}_i = \{\mathbf{xe}_{ij} | j = 1, ..., N\}$ by $\mathbf{xe}_{ij} = Enc(\mathbf{we}_{ij})$, where $Enc()$ denotes a word embedding lookup table. Then the sequence of word embeddings are fed into the word-level Bi-LSTM to produce an output representation of the words $\mathbf{le}^w$:

$$\mathbf{le}^w_{ij} = \overleftrightarrow{\text{LSTM}}(\mathbf{xe}_{ij}), j \in [1, N] \,. \tag{4.1}$$

To utilize the multi-head attention mechanism to obtain $\mathbf{ae}^w_i = \{\mathbf{ae}^w_{i1}, ..., \mathbf{ae}^w_{iN}\}$ at word level, we define $Q_i = \mathbf{le}^w_i$, $K_i = V_i = \mathbf{xe}_i$,

$$\mathbf{ae}^w_i = \text{MultiHead}(Q_i, K_i, V_i) \,. \tag{4.2}$$

The concatenation of $\mathbf{le}^w_i$ and $\mathbf{ae}^w_i$ of the words are fed into a Bi-LSTM and the output is concatenated to obtain the local context representation $\mathbf{he}^{ws}_i$ for each sentence $\mathbf{s_i}$:

$$\begin{aligned} \mathbf{he}^w_{ij} = \overleftrightarrow{\text{LSTM}}(\left[\mathbf{le}^w_{ij}; \mathbf{ae}^w_{ij}\right]), j \in [1, N] \ , \\ \mathbf{he}^{ws}_i = [\mathbf{he}^w_{i1}, ..., \mathbf{he}^w_{iN}] \ . \end{aligned} \tag{4.3}$$

To further model the global context between sentences, we apply a similar structure at sentence level. $\mathbf{he}^{ws} = \{\mathbf{he}^{ws}_i | i = 1, ..., M\}$ are fed into the sentence-level Bi-LSTM to produce output representation of the sentences $\mathbf{le}^s$:

$$\mathbf{le}^s_i = \overleftrightarrow{\text{LSTM}}(\mathbf{he}^{ws}_i), i \in [1, M] \ . \tag{4.4}$$

To utilize the multi-head attention mechanism to obtain $\mathbf{ae}^s = \{\mathbf{ae}^s_1, ..., \mathbf{ae}^s_M\}$ at sentence level, we define $Q = \mathbf{le}^s$, $K = V = \mathbf{he}^{ws}$,

$$\mathbf{ae}^s = \text{MultiHead}(Q, K, V). \tag{4.5}$$

The concatenation of the Bi-LSTM output $\mathbf{le}^s$ and the multi-head attention output $\mathbf{ae}^s$ of the sentences are fed into a Bi-LSTM to obtain the final representations of sentences $\mathbf{he}^s = \{\mathbf{he}^s_1, ..., \mathbf{he}^s_M\}$:

$$\mathbf{he}^s_i = \overleftrightarrow{\text{LSTM}}([\mathbf{le}^s_i; \mathbf{ae}^s_i]), i \in [1, M] \,. \tag{4.6}$$

4.2.1.2 Sentence-Level Extraction

Similar to [11], we use an LSTM-based Pointer Network to decode the above sentence representations $\mathbf{he}^s = \{\mathbf{he}^s_1, ..., \mathbf{he}^s_M\}$ and extract sentences recurrently to form an extractive summary $\mathbf{S_E} = \{A_1, ..., A_k, ..., A_{L_E}\}$ with L_E sentences, where A_k denotes the k-th sentence extracted.

At the k-th time step, the pointer network receives the sentence representation of the previous extracted sentence and has hidden state de_k. It first obtains a context vector de'_k by attending to $\mathbf{he}^s$:

$$\begin{aligned} \mathbf{ue}^k_i &= v^T \tanh(W_1 \mathbf{he}^s_i + W_2 de_k), i \in (1, ..., M) \ , \\ \mathbf{ae}^k_i &= \text{softmax}(\mathbf{ue}^k_i), i \in (1, ..., M) \ , \\ de'_k &= \sum_{i=1}^{M} \mathbf{ae}^k_i \mathbf{he}^s_i \ , \end{aligned} \tag{4.7}$$

where v, W_1, W_2 are learnable parameters of the pointer network. Then it predicts the extraction probability $p(A_k)$ of a sentence:

$$\begin{aligned} de_k &\leftarrow [de_k, de'_k] \ , \\ \mathbf{ue}^k_i &= v^T \tanh(W_1 \mathbf{he}^s_i + W_2 de_k), i \in (1, ..., M) \ , \\ p(A_k | A_1, ..., A_{k-1}) &= \text{softmax}(\mathbf{ue}^k) \ . \end{aligned} \tag{4.8}$$

Decoding iterates until L_E sentences are selected to form S_E.

4.2.2 Compressor Agent

Given an extractive summary $\mathbf{S_E}$ consisting of a sequence of words $\mathbf{wc} = \{\mathbf{wc}_i | i = 1, ..., N\}$, the compressor agent aims to produce a compressive summary $\mathbf{S_C}$ by selecting L_C words from $\mathbf{S_E}$. As illustrated in Figure 4.3, it has a multi-head attentional Bi-LSTM model to learn the word representations. It uses a pointer network to extract words based on their representations.

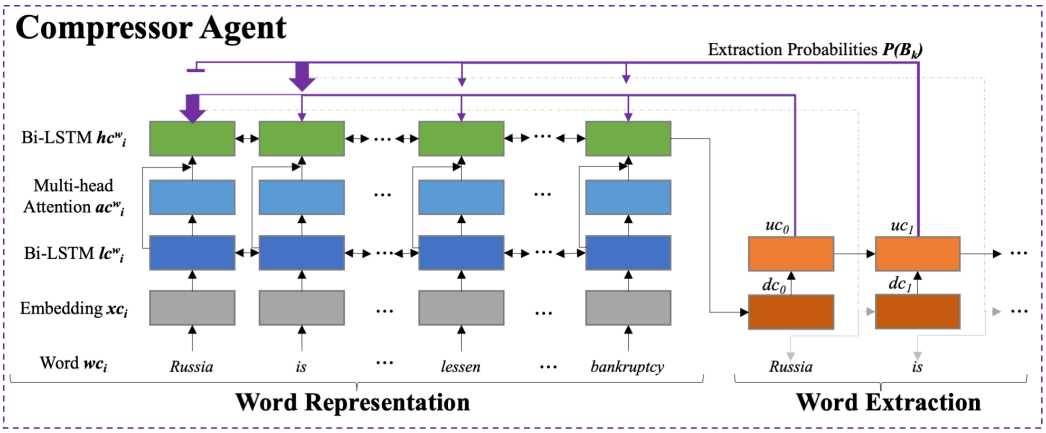

FIGURE 4.3. Illustration of the compressor agent.

4.2.2.1 Word Representation

Given a sequence of words $\mathbf{wc}$, we encode the words into word embeddings $\mathbf{xc} = \{\mathbf{xc}_i | i = 1, ..., N\}$ by $\mathbf{xc}_i = Enc(\mathbf{wc}_i)$. Then the sequence of word embeddings are fed into a Bi-LSTM to produce the words' output representation $\mathbf{lc}^w$:

$$\mathbf{lc}_i^w = \overleftrightarrow{\text{LSTM}}(\mathbf{xc}_i), i \in [1, N]\,. \tag{4.9}$$

To utilise the multi-head attention mechanism to obtain $\mathbf{ac}^w = \{\mathbf{ac}_1^w, ..., \mathbf{ac}_N^w\}$, we define $Q = \mathbf{lc}^w$, $K = V = \mathbf{xc}$,

$$\mathbf{ac}^w = \text{MultiHead}(Q, K, V). \tag{4.10}$$

The concatenation of $\mathbf{lc}^w$ and $\mathbf{ac}^w$ of the words are fed into a Bi-LSTM to obtain the representation $\mathbf{hc}_i^w$ for each word $\mathbf{wc_i}$:

$$\mathbf{hc}_i^w = \overleftrightarrow{\text{LSTM}}([\mathbf{lc}_i^w; \mathbf{ac}_i^w]), i \in [1, N]\,. \tag{4.11}$$

4.2.2.2 Word-Level Extraction

The word extractor of the compressor agent shares the same structure as that of the extractor agent's sentence extractor. To select the words based on the above word representations $\mathbf{hc}^w = \{\mathbf{hc}_1^w, ..., \mathbf{hc}_N^w\}$, the word extractor decodes and extracts words recurrently to produce $\{B_1, ..., B_k, ..., B_{L_C}\}$, where B_k denotes the word extracted at the k-th time step. The selected

words are reordered by their locations in the input document and assembled to form the compressive summary $\mathbf{S_C}$.

4.2.3 Reward in Reinforcement Learning

We use the compressive summary $\mathbf{S_C}$ to compute the reward of reinforcement learning and denote Reward$(\mathbf{D}, \mathbf{S_C})$ as Reward$(\mathbf{D}, \mathbf{S})$ for simplicity. Reward$(\mathbf{D}, \mathbf{S})$ is a weighted sum of the semantic coverage award Reward$_{\text{cov}}(\mathbf{D}, \mathbf{S})$ and the fluency reward Reward$_{\text{flu}}(\mathbf{S})$:

$$\begin{aligned} \text{Reward}(\mathbf{D}, \mathbf{S}) = w_{\text{cov}}&\text{Reward}_{\text{cov}}(\mathbf{D}, \mathbf{S}) \\ &+ w_{\text{flu}}\text{Reward}_{\text{flu}}(\mathbf{S}) \ , \end{aligned} \tag{4.12}$$

where w_{cov} and w_{flu} denote the weights of two rewards.

4.2.3.1 Semantic Coverage Reward

We compute Reward$_{\text{cov}}$ with the Wasserstein distance between the corresponding semantic distributions of the document $\mathbf{D}$ and the summary $\mathbf{S}$, which is the minimum cost required to transport the semantics from $\mathbf{D}$ to $\mathbf{S}$. We denote $\mathbf{D} = \{d_i | i = 1, ..., N\}$ to represent a document, where d_i indicates the count of the i-th token (i.e., word or phrase in a vocabulary of size N). Similarly, for a summary $\mathbf{S} = \{s_j | j = 1, ..., N\}$, s_j is respect to the count of the j-th token . The semantic distribution of a document is characterized in terms of normalised term frequency without the stopwords. The term frequency of the i-th token in the document $\mathbf{D}$ and the j-th token in the summary $\mathbf{S}$ are denoted as $\text{TF}_{\mathbf{D}}(i)$ and $\text{TF}_{\mathbf{S}}(j)$, respectively. By defining $\text{TF}_{\mathbf{D}} = \{\text{TF}_{\mathbf{D}}(i)\} \in \mathbf{R}^N$ and $\text{TF}_{\mathbf{S}} = \{\text{TF}_{\mathbf{S}}(j)\} \in \mathbf{R}^N$, we have the semantic distributions within $\mathbf{D}$ and $\mathbf{S}$ respectively.

The transportation cost matrix $\mathbf{C}$ is obtained by measuring the semantic similarity between each of the tokens. Given a pre-trained tokeniser and token embedding model with N tokens, define $\mathbf{v}_i$ to represent the feature embedding of the i-th token. Then the transport cost c_{ij} from the i-th to the j-th token is computed based on the cosine similarity: $c_{ij} = 1 - \frac{<\mathbf{v}_i, \mathbf{v}_j>}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}$. An optimal transport plan $\mathbf{T}^* = \{t^*_{i,j}\} \in \mathbf{R}^{N \times N}$ in pursuit of minimizing the transportation cost

can be obtained by solving the optimal transportation and resources allocation optimization problem [94]. Note that the transport plan can be used to interpret the transportation of tokens from document to summary, which brings interpretability to our URLComSum method.

Wasserstein distance measuring the distance between the two semantic distributions $\text{TF}_{\mathbf{D}}$ and $\text{TF}_{\mathbf{S}}$ with the optimal transport plan is computed by: $d_W(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}}|\mathbf{C}) = \sum_{i,j} t^*_{ij} c_{ij}$. $\text{Reward}_{\text{cov}}(\mathbf{D}, \mathbf{S})$ can be further defined as:

$$\text{Reward}_{\text{cov}}(\mathbf{D}, \mathbf{S}) = 1 - d_W(\text{TF}_{\mathbf{D}}, \text{TF}_{\mathbf{S}}|\mathbf{C}) \ . \tag{4.13}$$

4.2.3.2 Fluency Reward

We utilise Syntactic Log-Odds Ratio (SLOR) [91] to measure $\text{Reward}_{\text{flu}}(S)$, which is defined as: $\text{Reward}_{\text{flu}}(S) = \frac{1}{|S|}(\log(P_{LM}(S)) - \log(P_U(S)))$, where $P_{LM}(S)$ denotes the probability of the summary assigned by a pre-trained language model LM, $p_U(S) = \prod_{t \in S} P(t)$ denotes the unigram probability for rare word adjustment, and $|S|$ denotes the sentence length.

We use the Self-Critical Sequence Training (SCST) method [103], since this training algorithm has demonstrated promising results in text summarisation [92, 48]. For a given input document, the model produces two separate output summaries: the sampled summary $\mathbf{S}^s$, obtained by sampling the next pointer t_i from the probability distribution at each time step i, and the baseline summary $\hat{\mathbf{S}}$, obtained by always picking the most likely next pointer t at each i. The training objective is to minimise the following loss:

$$\begin{aligned} Loss = -(&\text{Reward}(\mathbf{D}, \mathbf{S}^s) - \text{Reward}(\mathbf{D}, \hat{\mathbf{S}})) \\ &\cdot \frac{1}{N} \sum_{i=1}^{N} \log p(t^s_i | t^s_1, ..., t^s_{i-1}, \mathbf{D}) \ , \end{aligned} \tag{4.14}$$

where N denotes the length of the pointer sequence, which is the number of extracted sentences for the extractor agent and the number of extracted words for the compressor agent.

Minimising the loss is equivalent to maximising the conditional likelihood of $\mathbf{S}^s$ if the sampled summary $\mathbf{S}^s$ outperforms the baseline summary $\hat{\mathbf{S}}$, i.e. $\text{Reward}(\mathbf{D}, \mathbf{S}^s) - \text{Reward}(\mathbf{D}, \hat{\mathbf{S}}) > 0$, thus increasing the expected reward of the model.

4.3 Experimental Results

4.3.1 Experimental Settings

We conducted comprehensive experiments on three widely used datasets: *Newsroom* [29], *CNN/DailyMail (CNN/DM)* [34], and *XSum* [80]. We set the LSTM hidden size to 150 and the number of recurrent layers to 3. We performed hyperparameter searching for w_{cov} and w_{flu} and decided to set $w_{\text{cov}} = 1$, $w_{\text{flu}} = 2$ in all our experiments since it provides more balanced results across the datasets. We trained the URLComSum with AdamW [63] with learning rate 0.01 with a batch size of 3. We obtained the word embedding from the pre-trained GloVe [93]. We used BERT for the pre-trained embedding models used for computing semantic coverage reward. We chose GPT2 for the trained language model used for computing the fluency reward due to strong representation capacity.

As shown in Table 4.1, we followed [75] to set $\mathbf{L_E}$ for Newsroom and [139] to set $\mathbf{L_E}$ for CNN/DM and XSum. We also followed their protocols to set $\mathbf{L_C}$ by matching the average number of words in summaries.

Dataset	**Newsroom**	**CNN/DM**	**XSum**
#Sentences in Doc.	27	39	19
#Tokens in Doc.	659	766	367
$\mathbf{L_E}$	2	3	2
$\mathbf{L_C}$	26	58	24
Train	995,041	287,113	204,045
Test	108,862	11,490	11,334

TABLE 4.1. Overview of the three datasets. #Sentences in Doc. and #Tokens in Doc. denote the average number of sentences and words in the documents respectively. $\mathbf{L_E}$ denotes the number of sentences to be selected by the extractor agent. $\mathbf{L_C}$ denotes the number of words to be selected by the compressor agent. Train and Test denote the size of train and test sets.

We compare our model with existing compressive methods which are all supervised, including *LATENTCOM* [135], *EXCONSUMM* [75], *JECS* [122], *CUPS* [13]. Since our method is unsupervised, we also compare it with unsupervised extractive and abstractive methods, including *TextRank* [77], *PacSum* [138], *PMI* [87], and *SumLoop* [48]. To better evaluate

Method	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	33.9	23.2	30.7
LEAD-WORD	34.9	23.1	30.7
Supervised Methods			
EXCONSUMM (Ext.)*	31.9	16.3	26.9
EXCONSUMM (Ext.+Com.)*	25.5	11.0	21.1
Unsupervised Methods			
SumLoop (Abs.)	27.0	9.6	26.4
TextRank (Ext.)	24.5	10.1	20.1
URLComSum (Ext.)	33.9	**23.2**	30.0
URLComSum (Ext.+Com.)	**34.6**	22.9	**30.5**

TABLE 4.2. Comparisons on the **Newsroom** test set. The symbol * indicates that the model is not directly comparable to ours as it is based on a subset (the "Mixed") of the dataset.

Method	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	40.0	17.5	32.9
LEAD-WORD	39.7	16.6	32.5
Supervised Methods			
LATENTCOM (Ext.)	41.1	18.8	37.5
LATENTCOM (Ext.+Com.)	36.7	15.4	34.3
JECS (Ext.)	40.7	18.0	36.8
JECS (Ext.+Com.)	41.7	18.5	37.9
EXCONSUMM (Ext.)	41.7	18.6	37.8
EXCONSUMM (Ext.+Com.)	40.9	18.0	37.4
CUPS (Ext.)	43.7	20.6	40.0
CUPS (Ext.+Com.)	44.0	20.6	40.4
Unsupervised Methods			
SumLoop (Abs.)	37.7	14.8	**34.7**
TextRank (Ext.)	34.1	12.8	22.5
PacSum (Ext.)	**40.3**	**17.6**	24.9
PMI (Ext.)	36.7	14.5	23.3
URLComSum (Ext.)	40.0	17.5	32.9
URLComSum (Ext.+Com.)	39.3	16.0	32.2

TABLE 4.3. Comparisons between our URLComSum and the state-of-the-art methods on the **CNN/DM** test set. (Ext.), (Abs.), and (Com.) denote the method is extractive, abstractive, and compressive respectively.

compressive methods, we followed a similar concept as LEAD baseline [106] and created *LEAD-WORD* baseline which extracts the first several words of a document as a summary. The commonly used ROUGE metric [54] is adopted.

Method	ROUGE-1	ROUGE-2	ROUGE-L
LEAD	19.4	2.4	12.9
LEAD-WORD	18.3	1.9	12.8
Supervised Methods			
CUPS (Ext.)	24.2	5.0	18.3
CUPS (Ext.+Com.)	26.0	5.4	19.9
Unsupervised Methods			
TextRank (Ext.)	19.0	3.1	12.6
PacSum (Ext.)	**19.4**	2.7	12.4
PMI (Ext.)	19.1	**3.2**	12.5
URLComSum (Ext.)	**19.4**	2.4	**12.9**
URLComSum (Ext.+Com.)	18.0	1.8	12.7

TABLE 4.4. Comparisons between our URLComSum and the state-of-the-art methods on the **XSum** test set in terms of ROUGE F-score. URLComSum (Ext.) denotes the extractive summary produced by our extractor agent. URLComSum (Ext.+Com.) denotes the compressive summary produced further by our compressor agent.

4.3.2 Quantitative Analysis

The experimental results of URLComSum on different datasets are shown in Table 4.2, Table 4.3 and Table 4.4 in terms of ROUGE-1, ROUGE-2 and ROUGE-L F-scores. (Ext.), (Abs.), and (Com.) denote that the method is extractive, abstractive, and compressive, respectively. Note that on the three datasets, LEAD and LEAD-WORD baseline are considered strong baselines in the literature and sometimes perform better than the state-of-the-art supervised and unsupervised models. As also discussed in [106, 87], it could be due to the Inverted Pyramid writing structure [97] of news articles, in which important information is often located at the beginning of an article and a paragraph.

Our URLComSum method significantly outperforms all the unsupervised and supervised ones on Newsroom. This demonstrates the effectiveness of our proposed method. Note that, unlike supervised EXCONSUMM, our reward strategy contributes to performance improvement when the compressor agent is utilised. For example, in terms of ROUGE-L, EXCONSUMM(Ext.+Com.) does not outperform EXCONSUMM(Ext.), while URLComSum(Ext.+Com.) outperforms URLComSum(Ext.). Similarly, our URLComSum method achieves the best performance among all the unsupervised methods on XSum, in terms of

ROUGE-1 and -L. URLComSum underperforms in ROUGE-2, which may be due to the trade-off between informativeness and fluency. The improvement on Newsroom is greater than those on CNN/DM and XSum, which could be because the larger size of Newsroom is more helpful for training our model.

Our URLComSum method achieves comparable performance with other unsupervised methods on CNN/DM. Note that URLComSum does not explicitly take position information into account while some extractive methods take advantage of the lead bias of CNN/DM, such as PacSum and LEAD. Nevertheless, we observe that URLComSum(Ext.) achieves the same result as LEAD . Even though URLComSum is unsupervised, eventually the extractor agent learns to select the first few sentences of the documents, which follows the principle of the aforementioned Inverted Pyramid writing structure.

4.3.3 Ablation Studies

Effect of Compression. We observed that the extractive and compressive methods usually obtain better results than the abstractive ones in terms of ROUGE scores on CNN/DM and Newsroom, and vice versa on XSum. It may be that CNN/DM and Newsroom contain summaries that are usually more extractive, whereas XSum's summaries are highly abstractive. We noticed that URLComSum(Ext.+Com.) generally achieves higher ROUGE-1 and -L scores than its extractive version on Newsroom. Meanwhile, on CNN/DM and XSum, the compressive version has slightly lower ROUGE scores than the extractive version. We observe similar behaviour in the literature of compressive summarisation, which may be that the sentences of news articles have dense information and compression does not help much to further condense the content.

Effect of Transformer. Note that we investigated the popular transformer model [115] in our proposed framework to replace Bi-LSTM for learning the sentence and word representations. However, we noticed the transformer-based agents do not perform as well as the Bi-LSTM-based ones while training from scratch with the same training procedure. The difficulties of training a transformer model have also been discussed in [96, 58]. Besides, the commonly

used pre-trained transformer models, such as BERT [14] and BART [50], require high computational resources and usually use subword-based tokenizers. They are not suitable for URLComSum since our compressor agent points to words instead of subwords. Therefore, at this stage Bi-LSTM is a simpler and more efficient choice. Nevertheless, the transformer is a module that can be included in our framework and is worth further investigation in the future.

Comparison of Extraction, Abstraction and Compression Approaches. We observed that the extraction and compressive approaches usually obtain better results than the abstractive in terms of ROUGE scores on CNN/DM and Newsroom, and vice versa on XSum. It may be because CNN/DM and Newsroom contain summaries that are usually more extractive, whereas XSum's summaries are highly abstractive. Since the ROUGE metric reflects lexical matching only and overlooks the linguistic quality and factuality of the summary, it is difficult to conclude the superiority of one approach over the others solely based on the ROUGE scores. Automatic linguistic quality and factuality metrics would be essential to provide further insights and more meaningful comparisons.

4.3.4 Qualitative Analysis

In Figure 4.4, 4.5, 4.6, summaries produced by URLComSum are shown together with the reference summaries of the sample documents in the CNN/DM, XSum, and Newsroom datasets. This demonstrates that our proposed URLComSum method is able to identify salient sentences and words and produce reasonably fluent summaries even without supervision information.

The following shows the sample summaries generated by URLComSum on the CNN/DM, XSum, and Newsroom datasets. Sentences extracted by the URLComSum extractor agent are highlighted. Words selected by the URLComSum compressor agent are underlined in red. Our unsupervised method URLComSum can identify salient sentences and words to produce a summary with reasonable semantic coverage and fluency.

Figure 4.4 shows a sample summary produced by URLComSum on the CNN/DM dataset. The summary generated by URLComSum has ROUGE-1, ROUGE-2, and ROUGE-L F-Scores

of 68.8, 52.7, and 62.4 respectively, with semantic coverage reward 0.76 and fluency reward 0.64, while the reference summary has semantic coverage reward 0.80 and fluency reward 0.62.

Source Document:

Russia is considering bailing out Greece in exchange for the country's 'assets', it was reported last night. Alexis Tsipras, Greece's prime minister, will meet Vladimir Putin in Moscow today, amid reports that the Kremlin will offer controversial loans and discounts on supplies of natural gas in a bid to lessen its dependence on the West . The visit will raise fears the radical left government is looking east in search of alternative sources of finance as it bids to avoid bankruptcy. Scroll down for video . Alexis Tsipras, Greece's (...)

Reference Summary:

Alexis Tsipras, Greece's prime minister, will meet Vladimir Putin in Moscow . The meeting comes amid reports Russia is considering bailing out Greece . Reports Kremlin may offer loans and discounts on supplies of natural gas .

URLComSum:

Russia is considering bailing out Greece in exchange for the country ' s ' assets ' , it was reported last night . Alexis Tsipras , Greece ' s prime minister , will meet Vladimir Putin in Moscow today , amid reports that the Kremlin will offer controversial loans and discounts on supplies of natural gas in a bid its raise alternative as bids to avoid bankruptcy .

FIGURE 4.4. A sample summary produced by URLComSum on the CNN/DM dataset.

Figure 4.5 shows a sample summary produced by URLComSum on the XSum dataset. The summary generated by URLComSum has ROUGE-1, ROUGE-2, and ROUGE-L F-Scores of 38.1, 20.0, and 33.3 respectively, with semantic coverage reward 0.77 and fluency reward 0.56, while the reference summary has semantic coverage reward 0.73 and fluency reward 0.59.

Source Document:

Paul Robson is the second trader at the Dutch bank to plead guilty to trying to rig the Yen Libor rate and the first Briton to do so. Last year Rabobank paid $1bn (Â£597m) to US and European regulators for its part in the global rate-rigging scandal. Barclays Bank, Royal Bank of Scotland and Lloyds Bank have all previously been fined for rate rigging. (...)

Reference Summary:

A former senior trader at Rabobank has pleaded guilty to interest rate rigging in the US.

URLComSum:

Paul Robson is the second trader at the Dutch bank to plead guilty to trying to rig the Yen Libor rate and global rate-rigging scandal .

FIGURE 4.5. A sample summary produced by URLComSum on the XSum dataset.

Figure 4.6 shows a sample summary produced by URLComSum on the Newsroom dataset. The summary generated by URLComSum has ROUGE-1, ROUGE-2, and ROUGE-L F-Scores of 76.6, 62.2, and 76.6 respectively, with semantic coverage reward 0.79 and fluency

reward 0.61, while the reference summary has semantic coverage reward 0.76 and fluency reward 0.65.

Source Document:

A man armed with a rifle has killed four people in a rampage in Girona province, north-east Spain, police say. The gunman walked into a bar in the town of Olot, 120km (70 miles) north of Barcelona, and shot two men - reportedly a father and son who were both construction workers. Minutes later, he went to a bank and killed two staff, police said. (...)

Reference Summary:

A man armed with a rifle kills four people in a shooting rampage in north-east Spain, police say.

URLComSum:

A man armed with a rifle has killed four people in a rampage in Girona province , north-east Spain , police say . The gunman walked into a bar in

FIGURE 4.6. A sample summary produced by URLComSum on the Newsroom dataset.

4.3.5 Interpretable Visualisation of Semantic Coverage

URLComSum is able to provide an interpretable visualisation of the semantic coverage on the summarisation results through the transportation matrix. Figure 4.7 illustrates the transport plan heatmap, which associated with a resulting summary is illustrated. A heatmap indicates the transportation of semantic contents between tokens in the document and its resulting summary. The higher the intensity, the more the semantic content of a particular document token is covered by a summary token. Red line highlights the transportation from the document to the summary of semantic content of token "country", which appears in both the document and the summary. Purple line highlights how the semantic content of token "debt", which appears in the document only but not the summary, are transported to token "bankruptcy" and "loans", which are semantically closer and have lower transport cost, and thus achieve a minimum transportation cost in the OT plan.

4.4 Conclusion

In this chapter, we have presented URLComSum, the first unsupervised and an efficient method for compressive text summarisation. Our model consists of dual agents: an extractor agent and a compressor agent. The extractor agent first chooses salient sentences from a

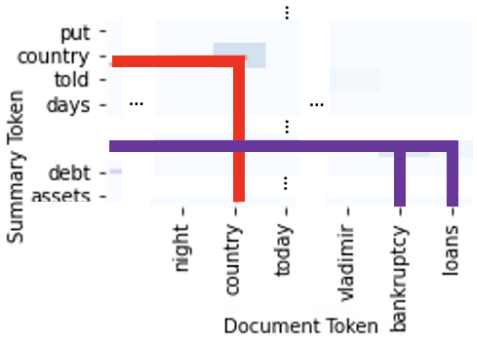

FIGURE 4.7. Interpretable visualisation of the OT plan.

document, and the compressor agent further select salient words from these extracted sentences to form a summary. To achieve unsupervised training of the extractor and compressor agents, we devise a reinforcement learning strategy to simulate human judgement on summary quality and optimize the summary's semantic coverage and fluency reward. Comprehensive experiments on three widely used benchmark datasets demonstrate the effectiveness of our proposed URLComSum and the great potential of unsupervised compressive summarisation. Our method provides interpretability of semantic coverage of summarisation results.

CHAPTER 5

TLDW: Extreme Multimodal Summarisation of News Videos

In this chapter, we introduce a new task, *eXtreme Multimodal Summarisation with Multimodal Output (XMSMO)* for the scenario of TL;DW - *Too Long; Didn't Watch*, akin to TL;DR. XMSMO aims to summarise a video-document pair into a summary with an extremely short length, which consists of one cover frame as the visual summary and one sentence as the textual summary. We propose a novel *u*nsupervised Hierarchical Optimal Transport Network (HOT-Net) consisting of three components: hierarchical multimodal encoders, hierarchical multimodal fusion decoders, and optimal transport solvers. Our method is trained, without using reference summaries, by optimising the visual and textual coverage from the perspectives of the distance between the semantic distributions under optimal transport plans. To facilitate the study on this task, we collect a large-scale dataset XMSMO-News by harvesting 4,891 video-document pairs. The experimental results show that our method achieves promising performance in terms of ROUGE and IoU metrics.

5.1 Introduction

Most of the existing MSMO methods are designed for short visual inputs, such as short videos and multiple images, without considering the summary length. Given the increasing pace of producing multimedia data and the subsequent challenge in keeping up with the explosive growth of such rich content, these existing methods may be sub-optimal to address the imminent issue of information overload of multimedia data.

In this chapter, we introduce a new task, *eXtreme Multimodal Summarisation with Multimodal Output (XMSMO)*, for the scenario TLDW which stands for *Too Long; Didn't Watch)*. As

FIGURE 5.1. Illustration of our newly proposed task XMSMO.

shown in Figure 5.1, XMSMO aims to summarise a pair of a video and its corresponding document into a multimodal summary with an extremely short length. That is, an extreme multimodal summary consists of one cover frame as the visual summary and one sentence as the textual summary. To solve this new task, we propose a novel unsupervised Hierarchical Optimal Transport Network (HOT-Net) architecture including three components, the hierarchical multimodal encoders, the hierarchical multimodal (fusion-based) decoders and the optimal transport solvers. The hierarchial structure could improve abstraction at multiple levels and multiple modalities and integration across these levels and modalities.

Specifically, the hierarchical visual encoder formulates the representations of a video from three levels including frame-level, scene-level and video-level; the hierarchical textual encoder formulates the representations of a document from three-levels as well: word-level, sentence-level and document-level. Then, the hierarchical decoder formulates the cross-modal representations in a local-global manner and evaluates candidate cover frames and candidate words, which are used to form a visual summary and a compressive textual summary, respectively. Note that a compressive textual summary offers a balance between the conciseness issue of extractive summarisation and the factual hallucination issue of abstractive summarisation.

Finally, our optimal transport-based unsupervised training strategy is devised to mimic human judgment on the quality of an extreme multimodal summary in terms of visual and textual coverage. The coverage is measured by a Wasserstein distance with an optimal transport plan measuring the distance between the semantic distributions of the summary and the original content. Wasserstein distance is used to take the semantic relationship between tokens into account and bring interpretability into the summarisation process. In addition, textual fluency and cross-modal similarity are further considered, which can be important to obtain a high-quality multimodal summary.

Additionally, to facilitate the study on this new task XMSMO and evaluate our proposed HOT-Net, we built the first dataset of such kind, namely XMSMO-News, by harvesting 4,891 video-document pairs as input and cover frame-title pairs as multimodal summary output from the British Broadcasting Corporation (BBC) News Youtube channel from year 2013 to 2021.

In summary, the key contributions of this chapter are:

- We introduce a new task, eXtreme Multimodal Summarisation with Multiple Output (XMSMO) as TLDW, which stands for *Too Long; Didn't Watch*. It aims to summarise a video-document pair into an extreme multimodal summary (i.e., one cover frame as the visual summary and one sentence as the textual summary).
- We propose a novel unsupervised Hierarchical Optimal Transport Network (HOT-Net). The hierarchical encoding and decoding are conducted across both the visual and textual modalities, which improve abstraction at multiple levels and multiple modalities and integration across these levels and modalities. Optimal transport solvers are introduced to guide the summaries to maximise their semantic coverage.
- We devise a new unsupervised training strategy that mimics the human judgment of a multimodal summary's quality by minimising the quartet loss of visual coverage, textual coverage, textual fluency, and cross-modal consistency.
- We constructed a new large-scale dataset, XMSMO-News, for the research community to facilitate research in this new direction. Experimental results on this

dataset demonstrate that our method outperforms other baselines in terms of ROUGE and IoU metrics.

The remainder of this chapter is organised as follows. Section 5.2 describe the details of our proposed method. Section 5.3 presents comprehensive experiments to evaluate the effectiveness of our proposed method. Lastly, Section 5.4 concludes our study with discussions on our future work.

5.2 Proposed Method

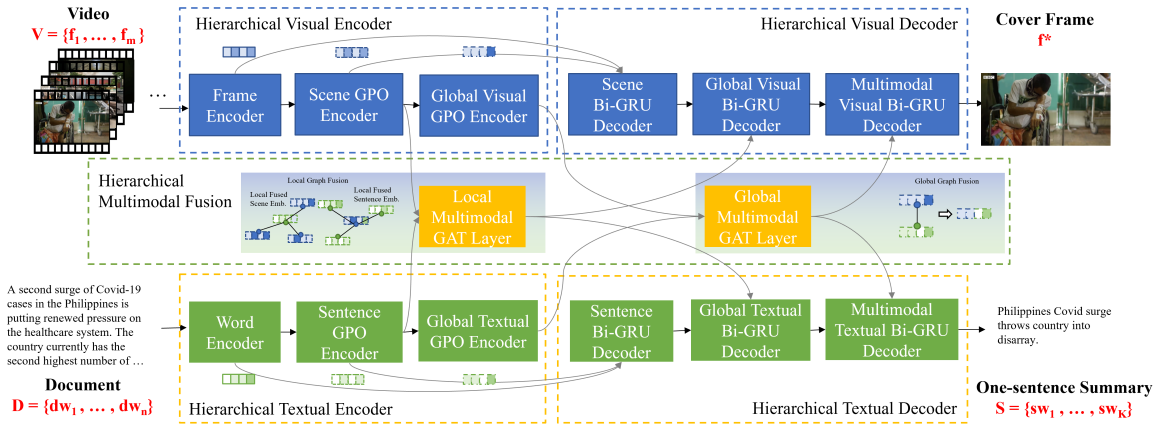

FIGURE 5.2. Illustration of the hierarchical multimodal encoder and hierarchical multimodal fusion decoder of our unsupervised Hierarchical Optimal Transport Network (HOT-Net) proposed for XMSMO.

As shown in Figure 5.2, our proposed eXtreme Multimodal Summarisation method, namely unsupervised Hierarchical Optimal Transport Network (HOT-Net), consists of three components, the hierarchical multimodal encoders, the hierarchical multimodal (fusion-based) decoders and the optimal transport solvers. Specifically, the hierarchical visual encoder formulates frame-level, scene-level and video-level representations of a video $\mathbf{V}$. The hierarchical textual encoder formulates word-level, sentence-level and document-level representations of a document $\mathbf{D}$. Then, the hierarchical visual decoder selects an optimal frame $\mathbf{f}^*$ as an extreme visual summary, and the hierarchical textual decoder produces an extreme textual summary

$\mathbf{s}^*$ based on the cross-modal guidance. Finally, the optimal transport solvers conduct unsupervised learning to optimise the encoders and the decoders in pursuit of the best semantic coverage of the obtained summaries.

5.2.1 Hierarchical Multimodal Encoders

5.2.1.1 Hierarchical Visual Encoder

Given an input video $\mathbf{V}$, it can be represented as a sequence of T frames, i.e. $\mathbf{V} = \{\mathbf{x}_i^{\text{frame}} | i = 1, ...T\}$. By grouping the consecutive frames with similar semantics, this input video $\mathbf{V}$ can be segmented into a sequence of T' scenes, i.e. $\mathbf{V} = \{\mathbf{x}_j^{\text{scene}} | j = 1, ..., T'\}$, where $\mathbf{x}_j^{\text{scene}}$ consists of the video frames from the i_{j_0}-th to the i_{j_1}-th frame, where j_0 indicates the start index of the frame and j_1 indicates the end index of the frame for the j-th scene in the video. The hierarchical visual encoder learns the scene-level and video-level representations based on $\mathbf{x}_i^{\text{frame}}$ and $\mathbf{x}_j^{\text{scene}}$, respectively.

To characterize a video frame $\mathbf{x}_i^{\text{frame}}$, a pre-trained neural network can be introduced. The CLIP model [101] is adopted in this study since it is the state-of-the-art multi-modal embedding model. For the sake of convenience, we use the the symbol $\mathbf{x}_i^{\text{frame}}$ to represent this pre-trained feature of the i-th frame. To further model the scene-level features, a pooling method is introduced, which is denoted as a function g^{scene}. In detail, for the j-th scene, its representation $\mathbf{x}_j^{\text{scene}}$ can be obtained by observing its associated frame-level features $\mathbf{x}_i^{\text{frame}}$, $i = i_{j_0}, ..., i_{j_1}$ as:

$$\mathbf{x}_j^{\text{scene}} = g^{\text{scene}}(\{\mathbf{x}_{i_{j_0}}^{\text{frame}}, ..., \mathbf{x}_{i_{j_1}}^{\text{frame}}\}). \tag{5.1}$$

Particularly, a generalized pooling operator (GPO) [8] is adopted as the pooling method in this study, since it is shown to be an effective and efficient pooling strategy for different features. With the scene-level features, a pooled global (i.e., video-level) representation can be derived as:

$$\mathbf{x}^{\text{video}} = g^{\text{video}}(\{\mathbf{x}_1^{\text{scene}}, ..., \mathbf{x}_{T'}^{\text{scene}}\}), \tag{5.2}$$

where g^{video} is a video-level pooling function based on a GPO operator.

5.2.1.2 Hierarchical Textual Encoder

An input document $\mathbf{D}$ can be viewed as a sequence consisting of U words as $\{\mathbf{x}_m^{\text{word}}|m = 1, ..., U\}$, or a sequence of U' sentences $\{\mathbf{x}_n^{\text{sentence}}|n = 1, ..., U'\}$. The n-th sentence consists of consecutive words in $\mathbf{D}$ from the m_{n_0}-th to the $m_{n,1}$-th word. Similar to the visual encoder, a hierarchical textual encoder is introduced to learn the sentence-level and the document-level representation.

A pre-trained CLIP model is introduced to formulate the word-level features, which is denoted as $\mathbf{x}_m^{\text{word}}$ for the m-th word. Next, a pooling mechanism g^{sentence} is adopted to formulate the sentence-level features. In detail, the n-th sentence-level features can be computed as:

$$\mathbf{x}_n^{\text{sentence}} = g^{\text{sentence}}(\{\mathbf{x}_{m_{n_0}}^{\text{word}}, ..., \mathbf{x}_{m_{n,1}}^{\text{word}}\}). \tag{5.3}$$

Finally, the global representation of the document $\mathbf{D}$ can be derived based on the sentence-level features:

$$\mathbf{x}^{\text{document}} = g^{\text{document}}(\{\mathbf{x}_1^{\text{sentence}}, ..., \mathbf{x}_{U'}^{\text{sentence}}\}), \tag{5.4}$$

where g^{document} is a document-level pooling function based on GPO.

5.2.2 Hierarchical Multimodal Fusion

To attend and fuse the representations from the visual and textual modalities, we adopt a graph-based attention mechanism (GAT) [116]. This multimodal fusion formulation helps easily extend the attention layer to future additional modalities, such as an audio modality. Each modality feature can be treated as a vertex feature of a graph. The relationships between modalities are formulated by graph convolution to attend over the other modalities, which then updates the representations of each modality. Particularly, a hierarchical local, which focuses between scene and sentence levels, and global, which focuses between video and document levels, observations are introduced by a graph fusion strategy.

For local multimodal fusion, the representations of the scenes $\mathbf{x}^{\text{scene}} = \{\mathbf{x}_1^{\text{scene}}, ..., \mathbf{x}_{T'}^{\text{scene}}\}$ and sentences $\mathbf{x}^{\text{sentence}} = \{\mathbf{x}_1^{\text{sentence}}, ..., \mathbf{x}_{U'}^{\text{sentence}}\}$ are fed into graph fusion modules $f_{\text{local}}^{\text{scene}}$ and

$f_{\text{local}}^{\text{sentence}}$. The resulted representation, which can be viewed as an information exchange between modalities, are fed into an average pooling operator g^{avg} to obtain the local multimodal context representations $\dot{\mathbf{x}}_j^{\text{scene}}$ and $\dot{\mathbf{x}}_n^{\text{sentence}}$:

$$\begin{aligned}\dot{\mathbf{x}}_j^{\text{scene}} =& g^{\text{avg}}([f_{\text{local}}^{\text{scene}}(\mathbf{x}_j^{\text{scene}}; \mathbf{x}_1^{\text{sentence}}), ..., \\ & f_{\text{local}}^{\text{scene}}(\mathbf{x}_j^{\text{scene}}; \mathbf{x}_{U'}^{\text{sentence}})]),\end{aligned} \tag{5.5}$$

$$\begin{aligned}\dot{\mathbf{x}}_n^{\text{sentence}} =& g^{\text{avg}}([f_{\text{local}}^{\text{sentence}}(\mathbf{x}_n^{\text{sentence}}; \mathbf{x}_1^{\text{scene}}), ..., \\ & f_{\text{local}}^{\text{sentence}}(\mathbf{x}_n^{\text{sentence}}; \mathbf{x}_{T'}^{\text{scene}})]).\end{aligned} \tag{5.6}$$

For global multimodal fusion, the global representations of the document $\mathbf{x}^{\text{document}}$ and video $\mathbf{x}^{\text{video}}$ are fed into a graph fusion module f_{global}:

$$\dot{\mathbf{x}} = g^{\text{avg}}(f_{\text{global}}(\left[\mathbf{x}^{\text{video}}; \mathbf{x}^{\text{document}}\right])). \tag{5.7}$$

5.2.3 Hierarchical Multimodal Decoders

5.2.3.1 Visual Decoder

Our visual decoder consists of three stages: 1) scene-guided frame decoding, 2) video-guided frame decoding, and 3) cross-modality-guided frame decoding. It aims to evaluate the probability of a particular frame being a cover frame.

To produce a scene-aware decoding outcome of evaluating each frame, a scene-guided visual decoder h^{scene} derives a latent decoding $\mathbf{y}_j^{\text{scene}}$ for frames from i_{j_0} to i_{j_1}, $j = 1, ..., T'$, as follows:

$$\begin{aligned}\mathbf{y}_j^{\text{scene}} &= \{\mathbf{y}_{i_{j_0}}^{\text{scene-frame}}, ..., \mathbf{y}_{i_{j_1}}^{\text{scene-frame}}\} \\ &= h^{\text{scene}}(\{\mathbf{x}_{i_{j_0}}^{\text{frame}}, ..., \mathbf{x}_{i_{j_1}}^{\text{frame}}\}|\dot{\mathbf{x}}_j^{\text{scene}}),\end{aligned} \tag{5.8}$$

where h^{scene} is a bi-directional GRU [6] and $\dot{\mathbf{x}}_j^{\text{scene}}$ is a multimodal scene guidance, which can be viewed as a prior knowledge. Next, to produce a video-guided frame decoding outcome, we have:

$$\begin{aligned}\mathbf{y}^{\text{video}} &= \{\mathbf{y}_1^{\text{video-frame}}, ..., \mathbf{y}_T^{\text{video-frame}}\} \\ &= h^{\text{video}}(\{\mathbf{x}_{i_{j_0}}^{\text{frame}}, ..., \mathbf{x}_{i_{j_1}}^{\text{frame}}\}|\mathbf{x}^{\text{video}}),\end{aligned} \tag{5.9}$$

where h^{video} is a bi-directional GRU and $\mathbf{x}^{\text{video}}$ is a unimodal video guidance as a prior knowledge. Finally, to produce a global multimodal context-aware decoding, we adopt a Bi-GRU decoder $\dot{h}^{\text{video}}$ with the guidance of the cross-modal embedding $\dot{\mathbf{x}}$:

$$\begin{aligned}\dot{\mathbf{y}}^{\text{video}} &= \{\dot{\mathbf{y}}_1^{\text{video-frame}}, ..., \dot{\mathbf{y}}_T^{\text{video-frame}}\} \\ &= \dot{h}^{\text{video}}(\mathbf{y}_1^{\text{video-frame}}, ..., \mathbf{y}_T^{\text{video-frame}} | \dot{\mathbf{x}}).\end{aligned} \tag{5.10}$$

To this end, the optimal frame $\mathbf{f}^*$ is obtained with a frame-wise linear layer activated with a softmax function:

$$\mathbf{f}^* = \text{argmax}_t(\text{Linear}(\dot{\mathbf{y}}^{\text{video}})). \tag{5.11}$$

5.2.3.2 Textual Decoder

Similar to the visual decoder, the textual decoder also consists of three stages: 1) sentenced-guided word decoding, 2) document-guided word decoding, and 3) cross-modality-guided word decoding. It aims to evaluate the probability of a word being selected in a compressive summary.

To produce a sentence-aware decoding outcome, a sentence decoder h^{sentence} derives a latent decoding $\mathbf{y}_n^{\text{sentence}}$ for words from m_{n_0} to $m_{n,1}$, $n = 1, ..., U'$, where n_0 indicates the start index of the word and n_1 indicates the end index of the word for the n-th sentence in the document, as follows:

$$\begin{aligned}\mathbf{y}_n^{\text{sentence}} &= \{\mathbf{y}_{m_{n_0}}^{\text{sentence-word}}, ..., \mathbf{y}_{m_{n_1}}^{\text{sentence-word}}\} \\ &= h^{\text{sentence}}(\{\mathbf{x}_{m_{n_0}}^{\text{word}}, ..., \mathbf{x}_{m_{n,1}}^{\text{word}}\} | \dot{\mathbf{x}}_n^{\text{sentence}}),\end{aligned} \tag{5.12}$$

where h^{sentence} is a bi-directional GRU and $\dot{\mathbf{x}}_n^{\text{sentence}}$ is used as a prior knowledge for the multimodal sentence guidance. Then, to produce a document-level textual decoding, we have:

$$\begin{aligned}\mathbf{y}^{\text{document}} &= \{\mathbf{y}_1^{\text{document-word}}, ..., \mathbf{y}_U^{\text{document-word}}\} \\ &= h^{\text{document}}(\{\mathbf{x}_{m_{n_0}}^{\text{word}}, ..., \mathbf{x}_{m_{n,1}}^{\text{word}}\} | \mathbf{x}^{\text{document}}),\end{aligned} \tag{5.13}$$

where h^{document} is a bi-directional GRU and $\mathbf{x}_n^{\text{document}}$ is a unimodal document guidance. Finally, to produce a global cross-modal context-aware decoding for each word, a Bi-GRU decoder

$\dot{h}^{\text{document}}$ is adopted with the guidance of the global multimodal embedding $\dot{\mathbf{x}}$:

$$\begin{aligned}\dot{\mathbf{y}}^{\text{document}} &= \{\dot{\mathbf{y}}_1^{\text{document-word}}, ..., \dot{\mathbf{y}}_U^{\text{document-word}}\} \\ &= \dot{h}^{\text{document}}(\mathbf{y}_1^{\text{document-word}}, ..., \mathbf{y}_U^{\text{document-word}} | \dot{\mathbf{x}}).\end{aligned} \tag{5.14}$$

As a result, the optimal compressive summary $\mathbf{s}^*$ with length k is obtained by:

$$\mathbf{s}^* = \text{topk}(\text{Linear}(\dot{\mathbf{y}}^{\text{document}})). \tag{5.15}$$

Note that the selected k words are ranked in line with their scores obtained from the linear layer with a softmax activation. Thus, the sentence $\mathbf{s}^*$ can be constructed with these words and their orders.

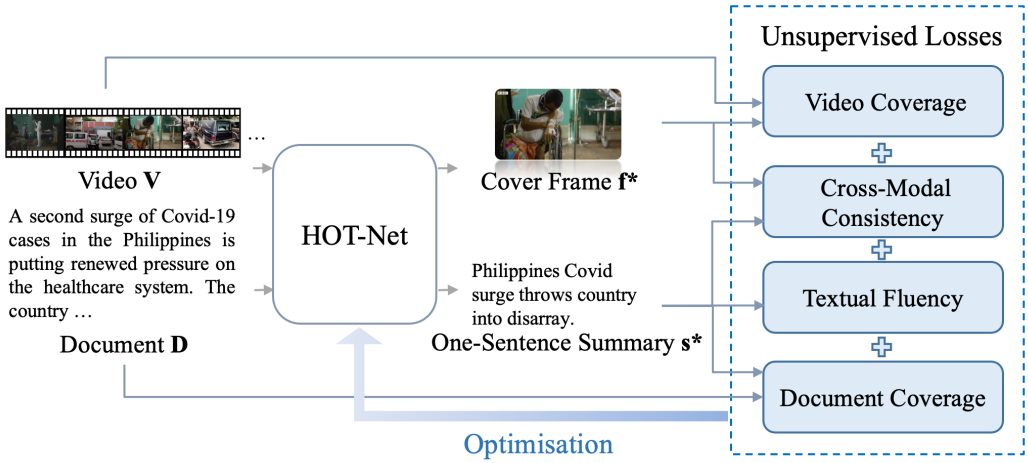

FIGURE 5.3. Optimal transport solver of HOT-Net for our unsupervised training strategy.

5.2.4 Optimal Transport-Guided Semantic Coverage

Our method is trained without reference summaries by mimicking the human judgment on the quality of a multimodal summary, which minimises a quartet loss of visual coverage, textual coverage, textual fluency, and cross-modal similarity. An overview of the training strategy is shown in Figure 5.3.

5.2.4.1 Optimal Transport-Guided Document Coverage

Intuitively, a high-quality summary is supposed to be close to the original document regarding their semantic distributions. Wasserstein distance is used to take the semantic relationship between tokens into account and bring interpretability into the summarisation process. We measure the Wasserstein distance [47] L_{document} between the document $\mathbf{D}$ and the selected sentence $\mathbf{s}^*$. It is the minimal cost required to transport the semantics from $\mathbf{s}^*$ to $\mathbf{D}$, measuring the semantic coverage of $\mathbf{s}^*$ on $\mathbf{D}$.

Given a dictionary, the number of the α-th token (i.e, a word in a dictionary) occurred in $\mathbf{D}$ can be counted as $P_{\mathbf{D}}(\alpha)$. As a result, the semantic distribution $\text{TF}_{\mathbf{D}}$ of the document $\mathbf{D}$ can be defined with the normalized term frequency of each token. In detail, for the α-th element of $\text{TF}_{\mathbf{D}}$, we have:

$$\text{TF}_{\mathbf{D}}(\alpha) = \frac{P_{\mathbf{D}}(\alpha)}{\sum_{\alpha'} P_{\mathbf{D}}(\alpha')}. \tag{5.16}$$

The semantic distribution $\text{TF}_{\mathbf{s}^*}$ of the selected sentence $\mathbf{s}^*$ can be derived in a similar manner. The normalized term frequency of the α-th token in $\mathbf{s}^*$ is:

$$\text{TF}_{\mathbf{s}^*}(\alpha) = \frac{P_{\mathbf{s}^*}(\alpha)}{\sum_{\alpha'} P_{\mathbf{s}^*}(\alpha')}. \tag{5.17}$$

Note that $\text{TF}_{\mathbf{D}}$ and $\text{TF}_{\mathbf{s}^*}$ have an equal total token quantities of 1 and can be completely transported from one to the other mathematically.

A transportation cost matrix $\mathbf{C} = (c_{\alpha\alpha'})$ is introduced to measure the semantic similarity between the tokens. Given a pre-trained tokeniser and token embedding model, define $\mathbf{u}_\alpha$ to represent the feature embedding of the α-th token. The transport cost $c_{\alpha\alpha'}$ from the α-th token to the α'-th one is computed based on the cosine similarity:

$$c_{\alpha\alpha'} = 1 - \frac{< \mathbf{u}_\alpha, \mathbf{u}_{\alpha'} >}{\|\mathbf{u}_\alpha\|_2 \, \|\mathbf{u}_{t\alpha'}\|_2}. \tag{5.18}$$

Note that the method to obtain token representations $\mathbf{u}_\alpha$ follows the same method that we formulate for word representations $\mathbf{x}^{\text{word}}_{(\cdot)}$ by a pre-trained model.

Then, an optimal transport plan matrix $\mathbf{T}^*(\mathbf{D}, \mathbf{s}^*) = (t^*_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*))$ in pursuit of minimizing the transportation cost can be obtained by solving the following optimization problem:

$$\begin{aligned} \mathbf{T}^*(\mathbf{D}, \mathbf{s}^*) &= \underset{\mathbf{T}(\mathbf{D}, \mathbf{s}^*)}{\operatorname{argmin}} \sum_{\alpha,\alpha'} t_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*) c_{\alpha\alpha}, \\ \text{s.t. } & \sum_{\alpha'} t_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*) = \mathrm{TF}_{\mathbf{D}}(\alpha), \\ & \sum_{\alpha=1} t_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*) t^{Doc}_{ij} = \mathrm{TF}_{\mathbf{s}^*}(\alpha'), \\ & t_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*) \geq 0, \\ & \forall \alpha, \alpha'. \end{aligned} \tag{5.19}$$

To this end, the Wasserstein distance can be defined as:

$$L_{\text{document}} = \sum_{\alpha,\alpha'} t^*_{\alpha\alpha'}(\mathbf{D}, \mathbf{s}^*) c_{\alpha\alpha'}, \tag{5.20}$$

which is associated with the optimal transport plan. By minimizing L_{document}, a high-quality summary sentence is expected to be obtained.

5.2.4.2 Optimal Transport-Guided Video Coverage

In parallel, a good cover frame is supposed to be close to the original video regarding their perceptual similarity. We measure the loss of visual coverage by computing the Wasserstein distance L_{video} between the corresponding colour signatures of the mean of video frames in $\mathbf{V}$ and the cover frame $\mathbf{f}^*$. It can be viewed as the minimal cost required to transport the semantics from $\mathbf{f}^*$ to $\mathbf{V}$.

By denoting $\bar{\mathbf{f}}$ as the mean of the video frames in $\mathbf{V}$, we define $\bar{\mathbf{r}}$ and $\mathbf{r}^*$ as the colour signatures of $\bar{\mathbf{f}}$ and $\mathbf{f}^*$, respectively. In detail, we have:

$$\begin{aligned} \bar{\mathbf{r}} &= \left\{(\bar{\mu}_1, \bar{\tau}_1), ..., (\bar{\mu}_{\bar{n}}, \bar{\tau}_{\bar{n}})\right\}, \\ \mathbf{r}^* &= \left\{(\mu^*_1, \tau^*_1), ..., (\mu^*_{m^*}, \tau^*_{m^*})\right\}, \end{aligned} \tag{5.21}$$

where $\bar{\mu}_i$ and μ^*_j are the points in the colour space, and $\bar{\tau}_i$ and τ^*_j are the corresponding weights of the points.

An optimal transport plan matrix $\mathbf{T}^*(\mathbf{V}, \mathbf{f}^*) = (t^*_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*)) \in \mathbf{R}^{\bar{m} \times m^*}$ in pursuit of minimizing the transportation cost between $\bar{\mathbf{r}}$ and $\mathbf{r}^*$ can be obtained by solving the following optimization problem:

$$\begin{aligned} \mathbf{T}^*(\mathbf{V}, \mathbf{f}^*) = \underset{\mathbf{T}(\mathbf{V}, \mathbf{f}^*)}{\operatorname{argmin}} \sum_{\beta, \beta'} t_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*) \left\| \bar{\mu}_\beta - \mu^*_{\beta'} \right\| \ , \\ \text{s.t.} \ \sum_{\beta'} t_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*) = \bar{\tau}_\beta, \\ \sum_{\beta} t_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*) = \tau^*_{\beta'} \ , \\ t_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*) \geq 0, \\ \forall \beta, \beta' \ , \end{aligned} \tag{5.22}$$

where $\mathbf{T}(\mathbf{V}, \mathbf{f}^*)$ is a transport plan. Then, a Wasserstein distance measuring the distance between the two colour signatures can be derived as:

$$L_{\text{video}} = t^*_{\beta\beta'}(\mathbf{V}, \mathbf{f}^*) \left\| \bar{\mu}_\beta - \mu^*_{\beta'} \right\|, \tag{5.23}$$

which is associated with the optimal transport plan. By minimizing L_{video}, a high-quality summary frame is expected to be the cover frame.

5.2.4.3 Textual Fluency

Inspired by [48], we adopt a pre-trained language model P_{LM} to measure the fluency of the textual summary L_{Fluency}. The loss can be defined as:

$$L_{\text{Fluency}} = P_{LM}(\mathbf{s}^*), \tag{5.24}$$

where P_{LM} computes the probability of $\mathbf{s}^*$ being a sentence.

5.2.4.4 Cross-modal Consistency

The semantic consistency should exist between the cover frame and the one-sentence summary. To formulate this, we measure the cross-modal similarity between the two embeddings of the cover frame $\mathbf{f}^*$ and the one-sentence summary $\mathbf{s}^*$. The loss can be defined based on a cosine

similarity:

$$L_{\text{cross-modal}} = 1 - cos(\mathbf{f}^*, \mathbf{s}^*). \tag{5.25}$$

In summary, four losses have been obtained to measure the summarisation quality: L_{document}, L_{video}, L_{fluency} and $L_{\text{cross-modal}}$. To this end, a loss function to optimize the proposed architecture can be formulated as follows:

$$\begin{aligned} L = \lambda_{\text{d}} L_{\text{document}} + \lambda_{\text{v}} L_{\text{video}} \\ + \lambda_{\text{f}} L_{\text{fluency}} + \lambda_{\text{c}} L_{\text{cross-modal}}, \end{aligned} \tag{5.26}$$

where λ_{d}, λ_{v}, λ_{f} and λ_{c} are the hyper-parameters controlling the weights of each loss term.

5.3 Experimental Results

5.3.1 Dataset

To the best of our knowledge, there is no existing large-scale dataset for XMSMO. Hence, we collected the first large-scale dataset of such kind, XMSMO-News, from the British Broadcasting Corporation (BBC) News Youtube channel [1]. We used the Pytube library to collect 4,891 quartets of video, document, cover frame, and one-sentence summary from the year 2013 to 2021. We used the video description as the document and video title as the one-sentence summary, as these visual and textual summaries were professionally created by the BBC. [2] We then split the quartets randomly into the train, validation, and test sets at a ratio 90:5:5. To facilitate future research that may utilise audio modality, we also collected the transcript of the video, which is automatically generated by Youtube. Six samples from XMSMO-News are shown in Figure 5.4. It shows that our dataset covers a wide variety of topics.

Table 5.1 shows the statistics and the comparison of our XMSMO-News dataset with other benchmarks on multimodal summarisation with multimodal output. The major differences

[1] https://www.youtube.com/c/BBCNews

[2] We removed the trailing promotional text from the video title and video description.

Title	Cover Frame	Description	Video	Transcript
Beirut: Why has there been crisis after crisis in Lebanon?		With Covid-19 infections on the rise, hospitals were already struggling to cope. Now, they are faced with treating thousands of injured people from Tuesday's explosion which shook Beirut. The country is also going through the worst economic crisis since the civil war, and tensions were already high with street demonstrations against the government. The BBC's Diplomatic Correspondent Caroline Hawley explains why Lebanon has had decades of unrest. Video by: Ameer Ahmed, Olivia Lace-Evans and Terry Saunders		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="0" dur="3.04">[Music]</text><text start="1.52" dur="2.399">think of lebanon and you might think of</text><text start="3.04" dur="3.359">its beauty</text>…
Malaysia's Anwar Ibrahim freed from jail after Mahathir election win.		Once seen as a future leader, he was jailed on sodomy and corruption charges after falling out with the government. Malaysia's new Prime Minister Mahathir Mohamad sought a pardon for Mr Anwar, which was granted on Wednesday morning. Mr Mahathir has promised to step aside for Mr Anwar to become prime minister within two years. The politician was jailed for a second time three years ago on what he said were trumped-up sodomy charges.		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="0" dur="4.77">after three prison sentences a total of</text><text start="2.52" dur="4.35">11 years in jail the Malaysian</text><text start="4.77" dur="4.74">opposition leader Anwar Ibrahim has been</text>…
Dozens killed in Myanmar protests as security forces fire on crowds.		The United Nations says at least 38 people have been killed during protests in Myanmar in the worst day of violence since the military coup last month. Security forces opened fire on large crowds in several cities across the country despite growing international condemnation. Huw Edwards presents BBC News at Ten reporting by south-east Asia correspondent Jonathan Head. #BBCNews		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="0.399" dur="4.081">the united nations says 38 people have</text><text start="2.399" dur="3.841">been killed during protests in myanmar</text>…
Fallen soldier's mother remembers last US military parade.		Donald Trump has vowed to hold a military parade in Washington that rivals France's Bastille Day. But Ileen Rollins, the mother of a fallen US soldier, recalls a different tone in the last US parade, in 1991, when the moment was all about commemoration. Video by Paul Blake World In Pictures https://www.youtube.com/playlist?list=PLS3XGiZxi7cBX37n4R0UGJN-TLiQOm7ZTP Big Hitters https://www.youtube.com/playlist?list=PLS3XGiZxi7cBUME-LUrFkDwFmiEc3jwMXP Just Good News https://www.youtube.com/playlist?list=PLS3XGiZxi7cBUsYo_P26cjihXLN-k3w246		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="2.24" dur="6.33">the stealth-bomber went over the trade</text><text start="5.73" dur="8.279">route and we've seen all the equipment</text><text start="8.57" dur="7.93">begins and the Humvees there was a lot</text><text start="14.009" dur="5.43">of marching men in their uniforms and</text><text start="16.5" dur="5.76">I'm sure I wasn't the only one that was</text>…
10 million people in UK given first vaccine dose but infections still alarmingly high		More than 10 million people across the UK have received at least one vaccination for coronavirus as part of the biggest programme of its kind in the history of the National Health Service. It's a major milestone in the race to control the faster-spreading variants of the virus. However medical experts say the UK 2019s infection levels are still alarmingly high 1d and many people are still dying from Covid-19. The pharma giant AstraZeneca and Oxford University say they hope to have created a next generation vaccine by the autumn, to protect against some new variants of the disease. Huw Edwards presents BBC News at Ten reporting by political editor Laura Kuenssberg, medical editor Fergus Walsh and health editor Hugh Pym. #BBCNews		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="0.08" dur="3.839">more than 10 million people across the</text><text start="1.92" dur="4.08">uk have now received at least one</text><text start="3.919" dur="4">vaccination for coronavirus</text><text start="6" dur="4.4">as part of the biggest program of its</text>…
America bids farewell to George HW Bush - BBC News		The Bush family has left the US Capitol. A military guard carried the former president's casket down the Capitol steps and placed it inside the hearse. They are making their way to the National Cathedral, an approximately 20-minute journey. The route is lined by mourners paying their respects. A memorial service for ex-President George HW Bush is held at Washington National Cathedral at 11:00 (1600 GMT) The nation's 41st president, who died on Friday aged 94, has lain in state since Monday in the rotunda of the US Capitol Bush's son, George W Bush, will give a eulogy. The three other living ex-US presidents will attend, and President Trump Britain's Prince Charles, German Chancellor Angela Merkel and Jordan's King Abdullah II are among foreign dignitaries His remains will be buried on Thursday at his presidential library in Texas next to wife, Barbara Bush, who died in April Please subscribe HERE http://bit.ly/1rbfUog		<?xml version="1.0" encoding="utf-8" ?><transcript><text start="0" dur="6.56">live pictures coming in as the final</text><text start="4.11" dur="5.19">rivals marked by the arrival of his son</text><text start="6.56" dur="5.89">the ex-president of course george w bush</text><text start="9.3" dur="7.55">who will be delivering one of the</text><text start="12.45" dur="6.989">eulogies at this service other</text><text start="16.85" dur="5.5">presidents including President Donald</text><text start="19.439" dur="5.43">Trump his predecessors Barack Obama Bill</text>

FIGURE 5.4. Some samples in our XMSMO-News dataset.

are regarding the input and output lengths: XMSMO-News has an average duration of 345.5 seconds, whereas VMSMO [52] has 60 seconds only.

Table 5.2 provides an analysis showing that the textual summary of XMSMO-News dataset is more challenging than that of MSMO dataset. We report the percentage of novel n-grams in the target gold summaries that do not appear in their source documents. There are 38.57% novel unigrams in the XMSMO-News reference summaries compared to 17.59% in MSMO dataset; the proportion of novel constructions grows for larger n-grams. This indicates that XMSMO-News textual summaries are more abstractive.

Previous study [72] shows that there may be factual hallucination issues in the titles of BBC news articles. In our future research, we plan to conduct a similar study to investigate if the

Dataset	XMSMO-News	VMSMO	MSMO
#Train/Val/Test	4382 / 252 / 257	180000 / 2460 / 2460	293965 / 10355 / 10262
Language	English	Chinese	English
Visual Input	Video	Video	Multiple unordered images
Textual Input	Document	Document	Document
Visual Output	Cover frame	Cover frame	One image
Textual Output	One-sentence	Arbitrary length	Multi-sentence
Frames/Video	8827.4	1500.0	6.6
Video Duration(s)	345.5	60.0	-
Tokens/Document	101.7	96.8	723.0
Tokens/Summary	12.4	11.2	70.0
Annotation	Full	Partial[1]	Partial[2]

[1] Not all ground-truth data is available;
[2] No visual ground-truth on training and validation splits.

TABLE 5.1. Comparison of XMSMO-News with existing MSMO benchmark datasets.

	XMSMO-News	MSMO
unigrams	**38.57**	17.59
bigrams	**78.24**	52.01
trigrams	**91.25**	69.49
4-grams	**95.58**	77.68

TABLE 5.2. The proportion of novel n-grams (%) in ground-truth summaries in XMSMO-News and MSMO datasets. Results are computed on the test set. We show that our XMSMO-News dataset is more abstractive and more challenging since the summary consists of more novel words.

textual summary in our XMSMO-News dataset may contain some hallucinated content that cannot be verified from the source video and document.

5.3.2 Implementation Details

We used the PyTorch library for the implementation of our method. We set the hidden size of GPO and GRU to 512. For the pre-trained CLIP model and the pre-trained token embedding model BERT (base version) used for computing the loss of textual coverage, we obtained them from HuggingFace [3]. To detect the scenes of a video, we utilised the PySceneDetect

[3]https://huggingface.co

library [4]. To compute the Wasserstein distances, we utilised the POT library [5] and the OpenCV library, respectively. For video preprocessing, we extracted one of every 360 frames to obtain 120 frames as candidate frames. All frames were resized to 640x360. We trained HOT-Net using AdamW [63] with a learning rate of 0.01 and a batch size of 3 for about 72 hours. All experiments were run on a GeForce GTX 1080Ti GPU card. For evaluation, we obtained our ROUGE scores by using the pyrouge package [6].

5.3.3 Baselines

To evaluate our proposed method HOT-Net, we compared it with the following categories of baseline methods. 1) Extreme multimodal summarisation method **PEGASUS-XSUM + CA-SUM** [131, 3], which is a combination of the state-of-the-art method of the extreme text summarisation task PEGASUS-XSUM [131] and that of the extreme video summarisation task CA-SUM [3], respectively; 2) Multi-modal summarisation with multimodal output approach includes: **VMSMO** [52], which is the state-of-the-art multimodal summarisation method utilising video and document as input, and zero-shot **CLIP** [101] method, which is based on the state-of-the-art multimodal embedding method CLIP with a fully connected layer for classification to perform multimodal summarisation; and 3) Unimodal extreme summarisation methods for reference include: **PG** [106], **ProphetNet** [99], and **PEGASUS-XSUM** [131], which are the state-of-the-art methods of extreme text summarisation, and **ARL**[2] and **CA-SUM** [3], which are the state-of-the-art method of extreme video summarisation. The baseline models PEGASUS-XSUM and CLIP were obtained from HuggingFace [120], PG was obtained from the Github[7], and ProphetNet [99], CA-SUM [3], and VMSMO [52] were obtained from the authors' implementations. CA-SUM was obtained from the author's Github [8]; VMSMO was obtained from the author's Github [9] with modifications on the latest libraries' update and bug fixing.

[4] http://scenedetect.com/en/latest/
[5] https://pythonot.github.io
[6] https://pypi.org/project/pyrouge/
[7] https://github.com/kukrishna/pointer-generator-pytorch-allennlp
[8] https://github.com/e-apostolidis/CA-SUM
[9] https://github.com/iriscxy/VMSMO

5.3.4 Quantitative Analysis

For the quantitative evaluation of a textual summary, we followed the same evaluation protocol as the baseline methods [131, 52] and adopt the commonly used ROUGE metric [54] for text summarisation. For the visual summary, the commonly used Intersection over Union (IoU) [107] and frame accuracy [76] metrics for video summarisation are adopted.

The ROUGE metric evaluates the content consistency between a generated summary and a reference summary. In detail, the ROUGE-n F-scores calculates the number of overlapping n-grams between a generated summary and a reference summary. The ROUGE-L F-score considers the longest common subsequence between a generated summary and a reference summary.

$$\text{ROUGE-n} = \frac{\text{\# overlapping n-grams}}{\text{\# n-grams in ground-truth summary}} \tag{5.27}$$

IoU metric evaluates the high-level semantic information consistency by counting the number of overlap concepts between the ground-truth cover frame and the generated one.

$$IoU = \frac{\text{Number of overlapping concepts}}{\text{Total number of concepts}} \tag{5.28}$$

Frame accuracy metric is to compare lower-level visual features, the ground-truth cover frame and generated cover frame are considered to be matching when pixel-level Euclidean distance is smaller than a predefined threshold. [10]

$$Accuracy = \frac{\text{Matching cover frame}}{\text{Number of ground-truth cover frames}} \tag{5.29}$$

To evaluate the overall performance on both modalities, we compute the overall evaluation as:

$$0.5 \times \frac{\text{IoU}}{\text{Best IoU}} + 0.5 \times \frac{\text{ROUGE-L}}{\text{Best ROUGE-L}}, \tag{5.30}$$

[10] We followed [76] to set the predefined threshold to 0.6.

where the best IoU and the best ROUGE-L are the best scores among all the evaluated methods.

Method	**Textual Evaluation**			**Visual Evaluation**		**Overall Evaluation**
	ROUGE-1	ROUGE-2	ROUGE-L	Frame Accuracy	IoU	
Extreme Text Summarisation						
PG [106]	2.43	0.08	2.25	-	-	-
ProphetNet [99]	3.77	0.09	3.56	-	-	-
PEGASUS-XSUM [131]	4.36	**0.12**	4.00	-	-	-
Extreme Video Summarisation						
ARL [2]	-	-	-	0.59	0.68	-
CA-SUM [3]	-	-	-	0.57	0.69	-
Multimodal Summarisation with Multimodal Output						
VMSMO [52]	*Divergence*	*Divergence*	*Divergence*	0.57	0.69	0.49
CLIP [101]	4.14	0.08	3.80	0.54	0.63	0.89
Extreme Multimodal Summarisation						
PEGASUS-XSUM + CA-SUM	4.36	**0.12**	4.00	0.57	0.69	0.95
HOT-Net (Ours) visual only	-	-	-	**0.60**	0.68	-
HOT-Net (Ours) textual only	3.85	0.05	3.60	-	-	-
HOT-Net (Ours) w/o multimodal fusion	3.99	0.05	3.73	0.56	**0.70**	0.93
HOT-Net (Ours) w/o local-level multimodal fusion	4.45	0.06	4.16	0.59	**0.70**	0.98
HOT-Net (Ours) w/o global-level multimodal fusion	3.65	0.06	3.45	0.58	0.68	0.88
HOT-Net (Ours) w/o fluency loss	4.58	0.06	4.28	0.57	0.68	0.98
HOT-Net (Ours) w/o cross-modal loss	4.58	0.06	4.28	0.57	0.68	0.98
HOT-Net (Ours)	**4.64**	0.07	**4.33**	0.57	0.68	**0.99**

TABLE 5.3. Comparisons between our HOT-Net and the state-of-the-art summarisation methods on XMSMO-News. Our method outperforms the baseline models in terms of ROUGE-1 and ROUGE-L, which demonstrate the quality of the generated extreme textual summary, and achieves promising results in terms of frame accuracy and IoU, which demonstrate the quality of the generated extreme visual summary.

The experimental results of HOT-Net on XMSMO-News are shown in Table 5.3 including ROUGE-1, ROUGE-2. and ROUGE-L F-scores, and IoU. Our method outperforms the baseline models in terms of ROUGE-1 and ROUGE-L, which demonstrate the quality of the generated extreme textual summary, and achieves promising results in terms of frame accuracy and IoU, which demonstrate the quality of the generated extreme visual summary. HOT-Net underperforms in terms of ROUGE-2, which may be due to the trade-off between informativeness and fluency. PEGASUS-XSUM was trained on massive text corpora which may help improve the fluency of natural language generation. This trade-off is further discussed in the Qualitative Analysis section. Our work is the first study on this new topic and we expect the performance to improve over time.

The ROUGE metric reflects lexical matching only and often overlook the conciseness, linguistic quality and factuality of a summary, which are the key quality of a good summary. Since XMSMO-News extreme summaries are highly abstractive and very short (the summaries

of the extreme text summarisation dataset XSum [80] have 23 tokens on average; meanwhile, textual summaries of XMSMO-News have 12 tokens on average), conciseness, linguistic quality and factuality metrics would be essential to provide further insights. However, human evaluation is highly subjective, which is challenging to draw a meaningful comparison and conclusion, as pointed out in [36]. We advocate developing automatic metrics of conciseness, linguistic quality, and factuality for more meaningful evaluations in future research.

5.3.5 Ablation Study

To study the effect of the proposed mechanisms, we compare a number of different settings of our HOT-Net and the results can be found in Table 2. We first observe that multimodal learning improves the modelling by comparing it to the visual or textual-only method. Our fusion strategy is also important to obtain high-quality textual summaries. The local-and-global hierarchical mechanism improves the results of the textual summary. However, it does not have much impact on the results of the visual summary, which may be due to that the overall model architecture has achieved its best possible potential in terms of producing a visual summary. Additionally, the fluency loss and cross-modal loss improve the textual summary as well.

5.3.6 Qualitative Analysis

Figure 5.5 compares the summaries produced by HOT-Net and the baseline methods, and the reference summary of a sample in the XMSMO-News dataset. The example on the left hand side is about a US congressman who made an unusual appearance and flipped upside down. The example on the right hand side is about US President Donald Trump's UK visit. The example demonstrates that our proposed HOT-Net method produces factually correct and reasonably fluent extreme textual summary that captures the essence of the document even without supervision. In comparison, as highlighted in red colour, PEGASUS-XSUM produces a fluent but unfaithful summary with information that does not occur in the original document. Most of the methods agree on the choice of the cover frame, whilst ours and CA-SUM are

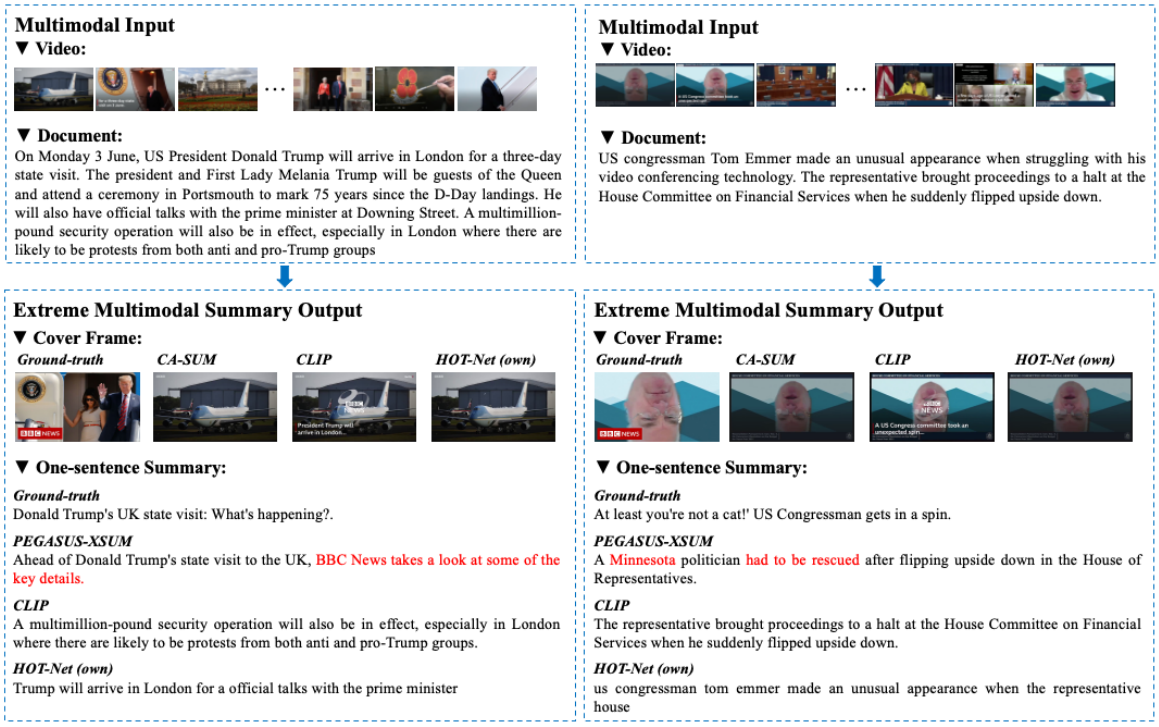

FIGURE 5.5. Example summaries generated by baseline methods and HOT-Net on XMSMO-News.

closer to the ground truth. For the second example, since the aeroplane appears repeatedly and occupies a comparatively large area on the frames, there is room for improvement to learn and identify the information which human considers to be *important*, such as a frame containing the face of the key human figure.

5.3.7 Interpretable Visualisation of Semantic Coverage

HOT-Net is able to provide an interpretable visualisation of the textual semantic coverage on the summarisation results. Figure 5.6 illustrates the transport plan heatmap, which indicates the transportation of semantic contents between tokens in the document and its resulting summary. The higher the colour intensity, the more the semantic content of a particular document token is covered by a summary token.

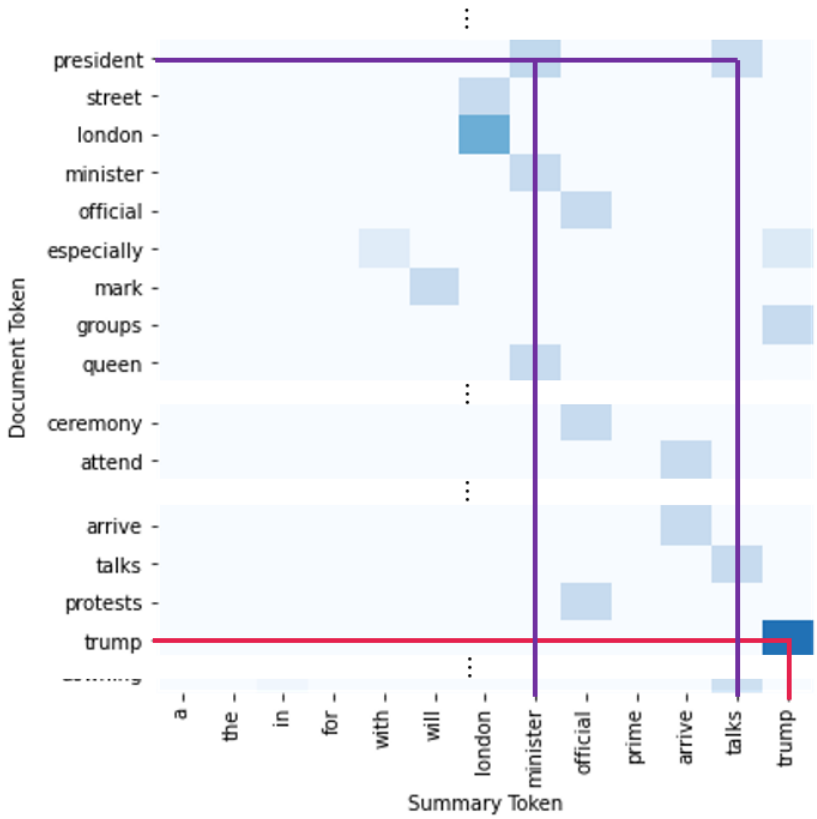

FIGURE 5.6. Interpretable visualisation of the OT plan from a source document to a resulting summary on the XMSMO-News dataset.

5.4 Conclusion

In this chapter, we have introduced a new task - eXtreme Multimodal Summarisation with Multimodal Output (XMSMO), which aims to summarise a video-document pair into an extreme multimodal summary, consisting of one cover frame as the visual summary and one sentence as the textual summary. We present a novel *unsupervised* deep learning architecture, which consists of three components: hierarchical multimodal encoders, hierarchical multimodal fusion decoders, and optimal transport solvers. To achieve unsupervised learning, besides the optimal transport-based semantic coverage guidance, textual fluency and cross-modal similarity are explored as well. In addition, we construct a new large-scale dataset XMSMO-News to facilitate research in this new direction. Experimental results demonstrate the effectiveness of our method.

According to the data analysis and experiments on our XMSMO-News dataset, our new task XMSMO has the following challenges: 1) Difficulty to identify the *most* salient textual and visual information. Since the summaries are extremely succinct, users would expect the summaries to include *only* the most important information. 2) Difficulty to ensure factuality and faithfulness. The extreme summaries is expected to give a only gaze to the full picture of the news event, it would be essential for users to trust that the summaries are representative and accurate descriptions of what happened without requiring further validation by digesting the details. 3) Difficulty to evaluate the model performance. Since the summaries are extremely short, the commonly used evaluation metrics of textual and visual summarisation, which are mainly designed to longer summaries and usually measure token and object overlapping, may penalise the false tokens and objects heavily and may not be a good fit to reflect the performance of the models. Developing automatic metrics that measure the qualities of the summaries, including informativeness, conciseness, linguistic and image quality, and factuality, would provide more meaningful evaluations.

In the future, we will explore the metric space to measure the optimal transport plan in a more efficient and effective manner. Moreover, we will explore improved ways to learn and identify the information that humans would consider to be *important*, such as a frame containing the face of a key character.

CHAPTER 6

TopicCAT: Unsupervised Topic-Guided Co-Attention Transformer for Extreme Multimodal Summarisation

---

In this chapter, we propose a novel Unsupervised Topic-guided Co-Attention Transformer (TopicCAT) based method to produce extreme multimodal summaries for video-document pairs. Specifically, there are two learning stages for a comprehensive multimodal understanding with a topic-based guidance: a unimodal learning stage and a cross-modal learning stage where a cross-modal topic model is devised to capture the overarching themes present in both documents and videos. To achieve unsupervised learning which does not require the resource-expensive collection of ground-truth multi-modal summaries, we propose an optimal transport-based optimisation scheme to evaluate summary coverage from a semantic distribution perspective at the topic level. Comprehensive experiments demonstrate the state-of-the-art performance of our proposed TopicCAT method on a new dataset containing 4,891 video-document pairs collected from British Broadcasting Corporation News Youtube with a BERTScore of 84.46 and an accuracy of 0.60.

6.1 Introduction

To tackle this emerging extreme multimodal summarisation task, we propose a novel Unsupervised Topic-guided Co-Attention Transformer (TopicCAT) method, which builds on the remarkable success of Transformers in sequential modelling [115]. Overall, TopicCAT consists of two learning stages for a comprehensive multimodal understanding with a topic-based guidance: a *unimodal learning stage* as Stage-I and a *cross-modal learning stage* as Stage-II. To capture the overarching themes present in both documents and videos effectively,

a cross-modal topic model is introduced, which is based on the embedding clustering of the two modalities jointly. In Stage-I, transformer-based encoder-decoder structures are devised for visual and textual learning individually to characterise unimodal patterns with their corresponding topic embeddings; while in Stage-II, a co-attention is adopted to obtain multimodal topic embeddings, which are treated as the multimodal topic guidance for the transformers in this stage to involve complementary multimodal contexts.

Our unsupervised training strategy is crafted to emulate human evaluation of the quality of an extreme multimodal summary by inspecting visual and textual topic coverage and cross-modal topic similarity. The approach avoids the need for resource-intensive collection of ground-truth summaries for supervision. Specifically, we devise novel loss functions based on optimal transport, assessing the differences in topic-level semantic distributions between: 1) video and cover frame representations; 2) document and one-sentence text summary representations; and 3) cover frame and one-sentence text summary representations. To address these three aspects, the network's weights are optimised to minimise three corresponding Wasserstein distances. informed by their optimal transport plans, which signify the minimal efforts to transform one distribution into another.

In summary, the key contributions of this chapter are as follows:

- A novel deep learning method - TopicCAT for the emerging extreme multimodal summarisation task based on Transformer neural networks, encompassing a two-stage unimodal and cross-modal learning strategy with topic guidance.
- A novel optimal transport guided unsupervised learning strategy is devised to optimise TopicCAT from the perspective of the similarity between semantic distributions of textual and visual topics.
- Comprehensive experiments and analysis are conducted on a new large-scale multimodal dataset demonstrate the effectiveness of the proposed TopicCAT method in achieving the state-of-the-art performance .

The remainder of this chapter is organised as follows. Section 6.2 describes the details of our proposed method. Section 6.3 presents comprehensive experiments to evaluate the

effectiveness of our proposed method. Lastly, Section 6.4 concludes our study with discussions on our future work.

6.2 Proposed Method

As shown in Figure 6.1, the proposed TopicCAT method consists of two learning stages with the topic guidance including a *unimodal learning stage* and a *cross-modal learning stage*. Our unsupervised learning method does not rely on ground-truth summaries as supervision for training, while an optimal transport based unsupervised learning strategy on the topic-level is proposed from a semantic distribution perspective.

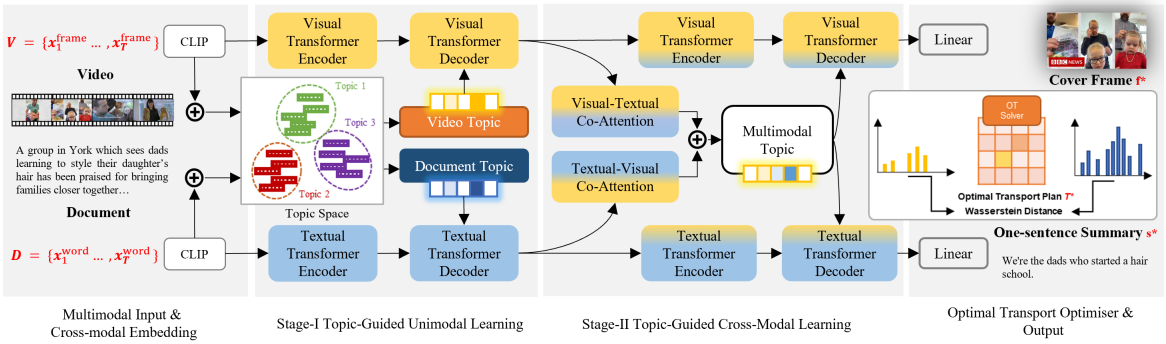

FIGURE 6.1. Illustration of the proposed Unsupervised Topic-Guided Co-Attention Transformer, namely TopicCAT, for extreme multimodal summarisation.

6.2.1 Visual and Textual Embeddings

To achieve contextualized and cross-modal consistency in embeddings for both documents and videos, the state-of-the-art CLIP model [101] is utilized to encode textual and visual data into a shared latent space. Specifically, an input video $\mathbf{V}$ can be represented by a sequence of frame embeddings, which is denoted as $\mathbf{V} = \{\mathbf{x}_i^{\text{frame}} | i = 1, ...T\}$, where T is the number of frames. Note that to maintain the sequential structure within the video sequence, the frame embeddings contain additional positional patterns by following the approach outlined in [115].

Moreover, the global representation of the video $\mathbf{V}$ can be derived with frame-level features:

$$\mathbf{x}^{\mathrm{v}} = g^{\mathrm{v}}(\{\mathbf{x}_1^{\mathrm{f}}, ..., \mathbf{x}_T^{\mathrm{f}}\}), \tag{6.1}$$

where g^{v} is a video-level pooling function. Specifically, g^{v} is based on a global average pooling which averages the frame features over time.

An input document $\mathbf{D}$ can be viewed as a sequence consisting of U words as $\{\mathbf{x}_m^{\mathrm{w}} | m = 1, ..., U\}$. Similarly, we use a pre-trained CLIP model with a position encoding to formulate the word-level features and $\mathbf{x}_m^{\mathrm{w}}$ denotes the embedding of the m-th word. The global representation of the document $\mathbf{D}$ can be derived based on the word-level features:

$$\mathbf{x}^{\mathrm{d}} = g^{\mathrm{d}}(\{\mathbf{x}_1^{\mathrm{w}}, ..., \mathbf{x}_U^{\mathrm{w}}\}), \tag{6.2}$$

where g^{d} is a document-level average pooling function.

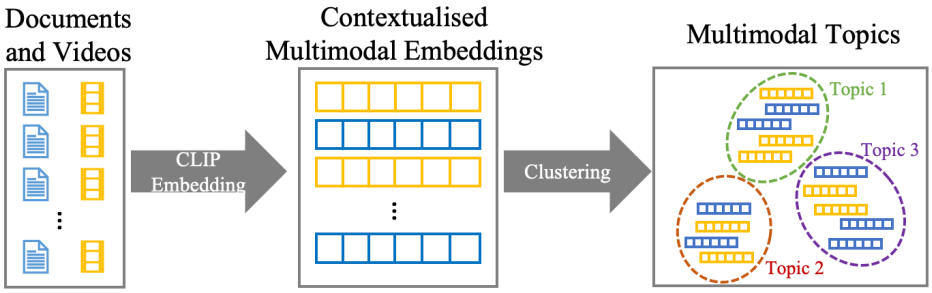

FIGURE 6.2. Illustration of cross-modal topic modelling to uncover the latent topics within multimodal inputs.

6.2.2 Crossmodal Topic Clustering

A crossmodal topic model aims to effectively encapsulate the overarching themes that present in both documents and videos. Our method for uncovering latent topics in document-video pairs encompasses a streamlined and efficient framework that clusters document and video embeddings jointly, as depicted in Figure 6.2. Specifically, we apply a clustering method to group similar documents and videos together into K clusters. Following [136], we opt for K-Means clustering owing to its efficiency and user-friendliness. The resulting clusters indicate the topics or themes presented in the document-video pairs, wherein each cluster

encompasses semantically related documents and videos. To obtain the topic vector $\mathbf{t}_{\mathbf{x}} \in \mathbb{R}^K$ for a given contextualised embedding $\mathbf{x}$, we transform it into the topic space through a topic function as follows:

$$\mathbf{t}_{x^{\mathrm{v}}} = \text{topic}(\mathbf{x}^{\mathrm{v}}), \mathbf{t}_{x^{\mathrm{d}}} = \text{topic}(\mathbf{x}^{\mathrm{d}}), \tag{6.3}$$

where each dimension of the obtained vector denotes the ℓ_2 distance to a cluster centre.

6.2.3 Stage-I: Topic-Guided Unimodal Learning

In this stage, summarisation is conducted on each modality individually with the corresponding topic guidance. Specifically, to determine whether a frame effectively represents the overall context of the video, we condition each frame based on the video topic vector. This is achieved by employing a Transformer-based encoder-decoder structure to obtain frame-level representations that are aware of the video's topic:

$$\begin{aligned}
\bar{\mathbf{x}}^{\mathrm{v},s_1} &= \{\bar{\mathbf{x}}_1^{\mathrm{f},s_1}, ..., \bar{\mathbf{x}}_T^{\mathrm{f}s_1}\} \\
&= \text{Enc}^{\mathrm{v},s_1}(\mathbf{x}^{\mathrm{v}}), \\
\mathbf{x}^{\mathrm{v},s_1} &= \{\mathbf{x}_1^{\mathrm{f},s_1}, ..., \mathbf{x}_T^{\mathrm{f},s_1}\} \\
&= \text{Dec}^{\mathrm{v},s_1}(\bar{\mathbf{x}}^{\mathrm{v},s_1} \oplus \mathbf{t}_{\mathbf{x}^{\mathrm{v}}}),
\end{aligned} \tag{6.4}$$

where $\oplus$ indicates a frame-wise concatenation with the video topic vector and the topic is introduced during the decoding procedure.

Likewise, to ascertain whether a word well reflects the overall context of the document, we condition each word on the document topic vector. This is achieved by utilizing the Transformer to generate a sequence of word representations that are aware of the document's topic:

$$\begin{aligned}
\bar{\mathbf{x}}^{\mathrm{d},s_1} &= \{\bar{\mathbf{x}}_1^{\mathrm{w},s_1}, ..., \bar{\mathbf{x}}_T^{\mathrm{w},s_1}\} \\
&= \text{Enc}^{\mathrm{d},s_1}(\mathbf{x}^{\mathrm{d}}), \\
\mathbf{x}^{\mathrm{d},s_1} &= \{\mathbf{x}_1^{\mathrm{w},s_1}, ..., \mathbf{x}_T^{\mathrm{w},s_1}\} \\
&= \text{Dec}^{\mathrm{d},s_1}(\bar{\mathbf{x}}^{\mathrm{d},s_1} \oplus \mathbf{t}_{\mathbf{x}^{\mathrm{d}}}).
\end{aligned} \tag{6.5}$$

6.2.4 Multimodal Co-Attention Transformer & Multimodal Topic Guidance

To better characterise the inter-dependency of cross-modal representations and develop an integrated understanding of both the document and the video, we introduce a multimodal co-attention mechanism. Specifically, we adapt the multi-head self-attention mechanism from Transformer Network, which takes into account only a single modality, to a co-attention mechanism that considers multiple modalities. The co-attention mechanism uses queries Q from one modality, and keys K and values V from another modality to generate features $\mathbf{x}^{\text{video, co}}$ and $\mathbf{x}^{\text{doc, co}}$ for one modality based on the other modality.

Mathematically, for frame embeddings in the video obtained in Stage-I, we have:

$$\mathbf{x}^{\text{v, co}} = \text{Linear}(Q + \text{Linear}(\text{MultiHead}(Q, K, V))), \tag{6.6}$$

where $Q = \mathbf{x}^{\text{v},s_1}$ and $K = V = \mathbf{x}^{\text{d},s_1}$. Similarly, for word embeddings in the document, we have:

$$\mathbf{x}^{\text{d, co}} = \text{Linear}(Q + \text{Linear}(\text{MultiHead}(Q, K, V))), \tag{6.7}$$

where $Q = \mathbf{x}^{\text{d},s_1}$, and $K = V = \mathbf{x}^{\text{v},s_1}$. Note that the weights of linear layers and multi-head co-attentions in Eq. (6.6) and Eq. (6.7) are different. For the sake of convenience, we do not differentiate them for the rest of the discussions in this chapter.

To this end, the resulting embeddings from the two co-attention mechanisms are fused to characterise the multimodal inputs comprehensively:

$$\mathbf{x}^{\text{d-v}} = \text{Linear}(\left[g^{\text{v}}(\mathbf{x}^{\text{v, co}}); g^{\text{d}}(\mathbf{x}^{\text{d, co}})\right]), \tag{6.8}$$

where g^{v} and g^{d} denote the global average pooling function. We then obtain the multimodal topic vector to represent the overall multimodal context with:

$$\mathbf{t}_{\mathbf{x}^{\text{d-v}}} = \text{topic}(\mathbf{x}^{\text{d-v}}). \tag{6.9}$$

6.2.5 Stage II: Topic-Guided Crossmodal Learning

With the multimodal topic guidance, which involves the context between different modalities, another stage of transformer based sequential learning is conducted. Specifically, the frame-level video representations in the second stage can be obtained as:

$$\begin{aligned}\bar{\mathbf{x}}^{\mathrm{v},s_2} &= \{\bar{\mathbf{x}}_1^{\mathrm{f},s_2}, ..., \bar{\mathbf{x}}_T^{\mathrm{f},s_2}\} \\ &= \mathrm{Enc}^{\mathrm{v},s_2}(\mathbf{x}^{\mathrm{v},s_1}); \\ \mathbf{x}^{\mathrm{v},s_2} &= \{\mathbf{x}_1^{\mathrm{f},s_2}, ..., \mathbf{x}_T^{\mathrm{f},s_2}\} \\ &= \mathrm{Dec}^{\mathrm{v},s_2}(\bar{\mathbf{x}}^{\mathrm{v},s_2} \oplus \mathbf{t}_{\mathbf{x}^{\mathrm{d\text{-}v}}}).\end{aligned} \tag{6.10}$$

To this end, the optimal frame $\mathbf{f}^*$ is obtained with 1) a frame-wise linear layer activated by a softmax function which estimates the probability that a frame to be a cover frame, and 2) the index of the frame that maximizes this probability is identified:

$$\mathbf{f}^* = \mathrm{argmax}(\mathrm{softmax}(\mathrm{Linear}(\mathbf{x}^{\mathrm{v},s_2}))). \tag{6.11}$$

Likewise, the word-level representations with multimodal topic guidance in the second stage can be formulated as:

$$\begin{aligned}\bar{\mathbf{x}}^{\mathrm{d},s_2} &= \{\bar{\mathbf{x}}_1^{\mathrm{w},s_2}, ..., \bar{\mathbf{x}}_T^{\mathrm{w},s_2}\} \\ &= \mathrm{Enc}^{\mathrm{d},,s_2}(\mathbf{x}^{\mathrm{d},s_1}); \\ \mathbf{x}^{\mathrm{d},s_2} &= \{\mathbf{x}_1^{\mathrm{w},s_2}, ..., \mathbf{x}_T^{\mathrm{w},s_2}\} \\ &= \mathrm{Dec}^{\mathrm{d},,s_2}(\bar{\mathbf{x}}^{\mathrm{d},s_2} \oplus \mathbf{t}_{\mathbf{x}^{\mathrm{d\text{-}v}}}).\end{aligned} \tag{6.12}$$

Next, an optimal compressive summary $\mathbf{s}^*$ with a budget k (i.e. summary length) can be obtained as:

$$\mathbf{s}^* = \text{top-k}(\mathrm{softmax}(\mathrm{Linear}(\mathbf{x}^{\mathrm{d},s_2}))). \tag{6.13}$$

Note that the selected k words are ranked in line with their scores obtained from the word-wise linear layer with a softmax activation. Thus, a summary sentence $\mathbf{s}^*$ can be constructed with these words in line with their orders.

6.2.6 Optimal Transport-Guided Unsupervised Training Strategy

Our training strategy aims to mimic human judgement on the quality of an extreme multimodal summary. Hence, we minimise a quartet loss function that considers 1) the coverage of visual topics $\mathcal{L}_v$, and 2) the coverage of textual topics from a semantic distribution perspective $\mathcal{L}_d$, 3) the consistency between the topics of different modalities $\mathcal{L}_{d-v}$, and 4) the fluency of textual summaries $\mathcal{L}_{fluency}$. Figure 6.3 illustrates our training strategy.

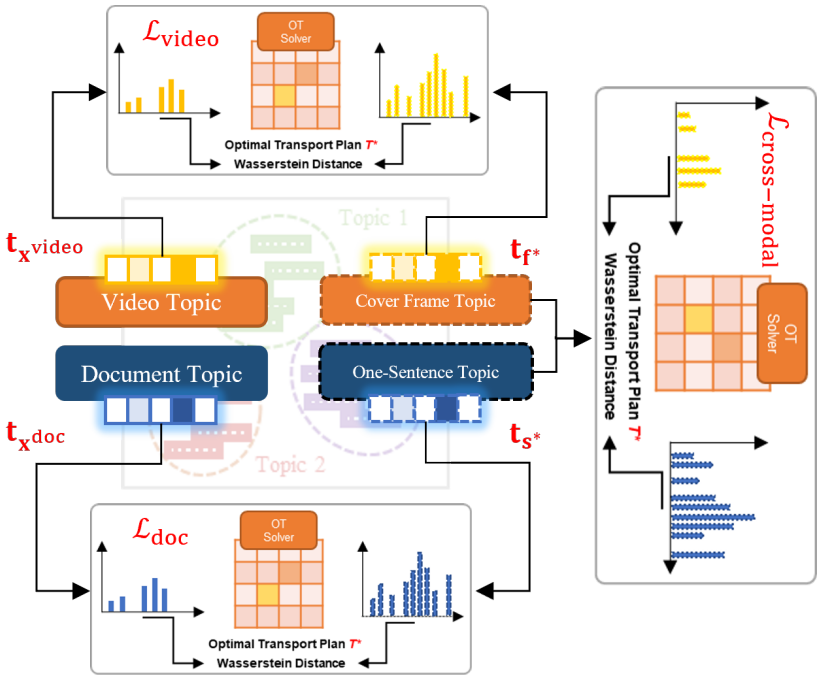

FIGURE 6.3. Illustration of the proposed optimal transport guided unsupervised training strategy.

6.2.6.1 Document Topic Coverage

Intuitively, a high-quality text summary regarding its semantic distribution is expected to closely reflect the semantic distribution of the original document. Specifically, we propose to measure the distribution difference at the topic level, which helps reduce the computational cost and observe the distributions from a global perspective.

Mathematically, the Wasserstein distance $\mathcal{L}_{Doc}$ is adopted [47] between the topic vector $\mathbf{t}_{x^{\mathrm{d}}}$ of the document $\mathbf{D}$ and the topic vector $\mathbf{t}_{\mathbf{s}}{}^* = \mathrm{topic}(\mathbf{s}^*)$ of the summary $\mathbf{s}^*$. It is the minimum

cost required to transport the topics from $\mathbf{s}^*$ to $\mathbf{D}$, measuring the topic coverage of $\mathbf{s}^*$ on $\mathbf{D}$. A transportation cost matrix $\mathbf{C} = \{c_{ij}\}$ is introduced to measure the transportation cost between the latent topics i and j. In this study, we assume a uniform transportation cost between the latent topics.

Then, an optimal transport plan $\mathbf{T}^*(\mathbf{D}, \mathbf{s}^*) = (t^*_{i,j}(\mathbf{D}, \mathbf{s}^*))$ in pursuit of minimising the transportation cost can be obtained by solving the following optimisation problem:

$$\begin{aligned} \mathbf{T}^*(\mathbf{D}, \mathbf{s}^*) &= \underset{\mathbf{T}(\mathbf{D},\mathbf{s}^*)}{\operatorname{argmin}} \sum_{i,j} t_{ij}(\mathbf{D}, \mathbf{s}^*) c_{ij}, \\ \text{s.t. } \sum_{j=1}^{N} t_{ij}(\mathbf{D}, \mathbf{s}^*) &= \mathbf{t}_{x^{\mathrm{d}}}, \sum_{i=1}^{N} t_{ij}(\mathbf{D}, \mathbf{s}^*) = \mathbf{t}_{\mathbf{s}^*}(j), \\ t_{ij}(\mathbf{D}, \mathbf{s}^*) &\geq 0, \forall i, j. \end{aligned} \tag{6.14}$$

To this end, the Wasserstein distance can be defined as:

$$\mathcal{L}_{\mathrm{d}} = \sum_{i,j} t^*_{ij}(\mathbf{D}, \mathbf{s}^*) c_{ij}, \tag{6.15}$$

which is associated with the optimal transport plan. By minimizing L_{Doc}, a high-quality textual summary $\mathbf{s}^*$ can be obtained.

6.2.6.2 Video Topic Coverage

Similarly, a good cover frame is expected to be close to the original video regarding the similarity of their topic distributions. Mathematically, we measure a loss based on video topic coverage by computing the Wasserstein distance $\mathcal{L}_{Video}$ between the topic vector $\mathbf{t}_{x^{\text{video}}}$ of the video $\mathbf{V}$ and the topic vector $\mathbf{t}_{\mathbf{f}^*}$ of the cover frame $\mathbf{f}^*$. It is the minimum cost to transport the topics from $\mathbf{f}^*$ to $\mathbf{V}$, and can be treated as the topic coverage of $\mathbf{V}$ by $\mathbf{f}^*$. Specifically, we

solve the following optimisation problem:

$$\begin{aligned} \mathbf{T}^*(\mathbf{V}, \mathbf{f}^*) &= \underset{\mathbf{T}(\mathbf{D}, \mathbf{f}^*)}{\operatorname{argmin}} \sum_{i,j} t_{ij}(\mathbf{V}, \mathbf{f}^*) c_{ij}, \\ \text{s.t. } \sum_{j=1}^{N} t_{ij}(\mathbf{V}, \mathbf{f}^*) &= \mathbf{t}_{x^{\text{v}}}, \sum_{i=1}^{N} t_{ij}(\mathbf{V}, \mathbf{f}^*) = \mathbf{t}_{\mathbf{f}^*}(j), \\ t_{ij}(\mathbf{V}, \mathbf{f}^*) &\geq 0, \forall i, j. \end{aligned} \tag{6.16}$$

To this end, the Wasserstein distance can be defined as:

$$\mathcal{L}_{\text{v}} = \sum_{i,j} t^*_{ij}(\mathbf{V}, \mathbf{f}^*) c_{ij}. \tag{6.17}$$

By minimizing L_v in terms of $\mathbf{f}^*$, $\mathbf{f}^*$ is expected to be well representative for the input video.

6.2.6.3 Cross-Modal Topic Consistency

The topic consistency should exist between the cover frame and the one-sentence summary. To formulate this aspect, we introduce a cross-modal topic consistency loss $\mathcal{L}_{Cross-modal}$ by computing the Wasserstein distance between the embeddings of the two topic vectors: the cover frame $\mathbf{f}^*$ and the one-sentence summary $\mathbf{s}^*$. In detail, we have:

$$\mathcal{L}_{\text{d-v}} = \sum_{i,j} t^*_{ij}(\mathbf{s}^*, \mathbf{f}^*) c_{ij}. \tag{6.18}$$

6.2.6.4 Textual Fluency

We utilise Syntactic Log-Odds Ratio (SLOR) [91] to measure L_{fluency}, which is defined as:

$$\mathcal{L}_{\text{fluency}}(S) = \frac{1}{|S|}(\log(P_{LM}(S)) - \log(P_U(S))) \,, \tag{6.19}$$

where $P_{LM}(S)$ represents the probability of the summary as assigned by a pre-trained language model LM, $p_U(S) = \prod_{t \in S} P(t)$ signifies the unigram probability used for rare word adjustment, and $|S|$ denotes the sentence length.

Overall, four loss terms have been derived to measure the quality of a multimodal summary: $\mathcal{L}_{\text{d}}$, $\mathcal{L}_{\text{v}}$, $\mathcal{L}_{\text{d-v}}$, and $\mathcal{L}_{\text{fluency}}$. Hence, the loss function to optimize the proposed TopicCAT model

can be formulated as follows:

$$\begin{aligned}\mathcal{L} &= \lambda_{\text{d}}\mathcal{L}_{\text{d}} + \lambda_{\text{v}}\mathcal{L}_{\text{v}} \\ &+\lambda_{\text{d-v}}\mathcal{L}_{\text{d-v}} + \lambda_{\text{fluency}}\mathcal{L}_{\text{fluency}},\end{aligned} \tag{6.20}$$

where λ_{d}, λ_{v}, $\lambda_{\text{d-v}}$, and λ_{fluency} are the hyperparameters controlling the weights of each term.

6.3 Experiments and Discussions

6.3.1 Dataset

We evaluated our proposed TopicCAT on a large-scale dataset for extreme multimodal summarisation. The new dataset was collected from the British Broadcasting Corporation (BBC) News Youtube channel[1], which has 4,891 quartets of video, document, cover frame, and one-sentence summary from Year 2013 to Year 2021. We utilized the video description as the document, while employing the video title as the one-sentence summary, as these visual and textual summaries were professionally created by the BBC. Subsequently, we split the quartets randomly into the training, validation, and test sets, adhering to a ratio 90:5:5. For video preprocessing, we selected one out of every 360 frames, resulting in 120 candidate frames. All frames were resized to 640x360 dimensions.

This dataset shares some similarities with the datasets proposed in [141] and [52] in terms of the input and output modalities, whilst there are two major differences between the dataset used in this study and those in previous works. The first difference pertains to the input and output lengths: the dataset used in this study has an average video duration of 345.5 seconds, while the videos in [52] generally last only 60 seconds, and the samples in [141] consist of a set of unordered images. A longer duration with more complex contents are more likely to be benefit from a topic modelling pipeline. The second distinction is related to summarization: the dataset in this study includes a one-sentence textual summary, whereas neither [52] nor [141] addresses the kind of extreme multimodal summarization that our work does. This

[1] https://www.youtube.com/c/BBCNews

allows us to evaluate our approach by taking a video and its associated document as input to produce a single image accompanied by a one-sentence text.

6.3.2 Implementation Details

We implemented our method using the PyTorch, setting the hidden size of the transformers to 512 and the number of heads to 4. Experiments were conducted with varying numbers of topic clusters K, ranging from 1 to 20. The optimal cluster number was determined to be 10, as identified by the elbow method. The pre-trained CLIP model for TopicCAT was obtained from HuggingFace [120].

Note that we found 5 epochs to be an optimal choice since the model reaches its best performance and does not improve further after this. Overall, TopicCAT has 466,663,683 number of learnable parameters and 1.8G memory footprint. Our experiments were conducted on a GeForce RTX 3090 GPU.

6.3.3 Baselines

We compared TopicCAT with the following categories of baselines. 1) MSMO approach includes: **VMSMO** [52], which is the state-of-the-art multimodal summarisation method utilising video and document as input, and zero-shot **CLIP** [101] method, which is based on the state-of-the-art multimodal embedding method CLIP with a fully connected layer for classification to perform multimodal summarisation; 2) Unimodal extreme summarisation methods for reference includes: **PG** [106], **PEGASUS-XSUM** [131] and **ProphetNet** [99], which are the state-of-the-art methods of extreme text summarisation, and **ARL**[2] and **CA-SUM** [3], which are the state-of-the-art method of extreme video summarisation. The CLIP model was obtained from HuggingFace [120]; PG was obtained from the Github [2]; ProphetNet [99], CA-SUM [3], and VMSMO [52] were obtained from the authors' implementations.

[2]https://github.com/kukrishna/pointer-generator-pytorch-allennlp

6.3.4 Quantitative Analysis

For the quantitative evaluation of a textual summary, BERTScore [133] and the commonly used ROUGE metric [54] are adopted , and the commonly used Intersection over Union (IoU) [107] and frame accuracy [76] metrics are adopted for video summarisation evaluation.

Method	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
Extreme Text Summarisation				
PG [106]	83.26	2.43	0.08	2.25
ProphetNet [99]	<u>85.53</u>	3.77	<u>0.09</u>	3.56
PEGASUS-XSUM [131]	**86.75**	4.36	**0.12**	4.00
Multimodal Summarisation with Multimodal Output				
VMSMO [52]	*Divergence*	*Divergence*	*Divergence*	*Divergence*
CLIP [101]	83.79	3.35	0.05	3.14
Extreme Multi-Modal Summarisation				
HOT-Net	84.00	**4.64**	0.07	**4.33**
TopicCAT (Ours)	84.46	<u>4.54</u>	0.08	<u>4.21</u>

TABLE 6.1. Comparisons between the textual evaluation of our TopicCAT and the state-of-the-art methods.

As shown in Table 6.1, our method's performance in text summarization compares favorably with most state-of-the-art methods for ROUGE-1 and ROUGE-L scores, highlighting its effectivenes. Note that TopicCAT falls short in BERTScore and ROUGE-2 compared to ProphetNet and PEGASUS-XSUM, which may be due to the trade-off between fluency and informativeness. Refer to Section 6.3.6 for further analysis on this. Note that ProphetNet is a large generative language model with a more complex architecture requiring significantly more training data. In comparison, our method has a more efficient architecture and training strategy and is able to achieve comparable results. Our work is the pioneering study on this new topic and we expect the performance to improve over time.

In terms of the visual summarisation, Table 6.2 lists the comparisons of the evaluation metrics between different methods. Our method achieves superior performance in terms of frame accuracy and competitive performance regarding IoU compared with baseline methods, which demonstrate the quality of the generated extreme visual summary.

Method	Frame Accuracy	IoU
Extreme Video Summarisation		
ARL [2]	0.59	0.68
CA-SUM [3]	0.57	**0.69**
Multimodal Summarisation with Multimodal Output		
VMSMO [52]	0.57	**0.69**
CLIP [101]	0.58	0.68
Extreme Multi-Modal Summarisation		
HOT-Net	0.57	0.68
TopicCAT (Ours)	**0.60**	**0.69**

TABLE 6.2. Comparisons between the visual evaluation of our TopicCAT and the state-of-the-art methods.

6.3.5 Ablation Study

To study the effect of the proposed mechanisms, a number of different settings of our TopicCAT are compared and the results are shown in Table 6.3.

6.3.5.1 Effects of Topic Guidance.

It is evident that intra- and inter-modal topic contexts enhance the modelling and outcomes of multimodal summarisation when contrasted with approaches that do not utilize this information. As detailed in Table 6.3, for intra-modal topic information, the one without visual topic information observes lower performance in terms of frame accuracy and IoU; and the one without textual topic information observes lower BERTScore performance and ROUGE-2, this may indicate it is less fluency, which is a crucial quality of extreme textual summary. For inter-topic information, the one without multimodal topic information shows diminished performance in all metrics.

Method	Textual Evaluation				Visual Evaluation	
	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	Frame Acc.	IoU
TopicCAT (Ours)	**84.46**	4.54	**0.08**	4.21	**0.60**	**0.69**
- multimodal topic	83.21	2.95	0.03	2.78	0.58	**0.69**
- visual topic	**84.46**	4.54	**0.08**	4.21	0.58	0.68
- textual topic	82.42	**4.66**	0.04	**4.30**	**0.60**	**0.69**

TABLE 6.3. Ablation studies on different settings of TopicCAT.

Training Strategy	Topic Coverage based	Semantic Coverage based
Training Time (hours)	17.68 (-7.3%)	19.07

TABLE 6.4. Comparison of training times with different strategies: topic coverage based vs full content coverage based.

6.3.5.2 Effects of Topic Coverage vs Semantic Coverage

In this study, we utilize the optimal transport-guided training loss, which focuses on the topic level coverage. Note that the Wasserstein distance and optimal transport problem can be formulated on the full contents (i.e., word distributions for documents and pixel distributions for videos) as well. We compared these two settings, where we observe that it has a negligible impact on quantitative performance but can reduce 7.3% training time as shown in Table 6.4.

6.3.6 Qualitative Analysis

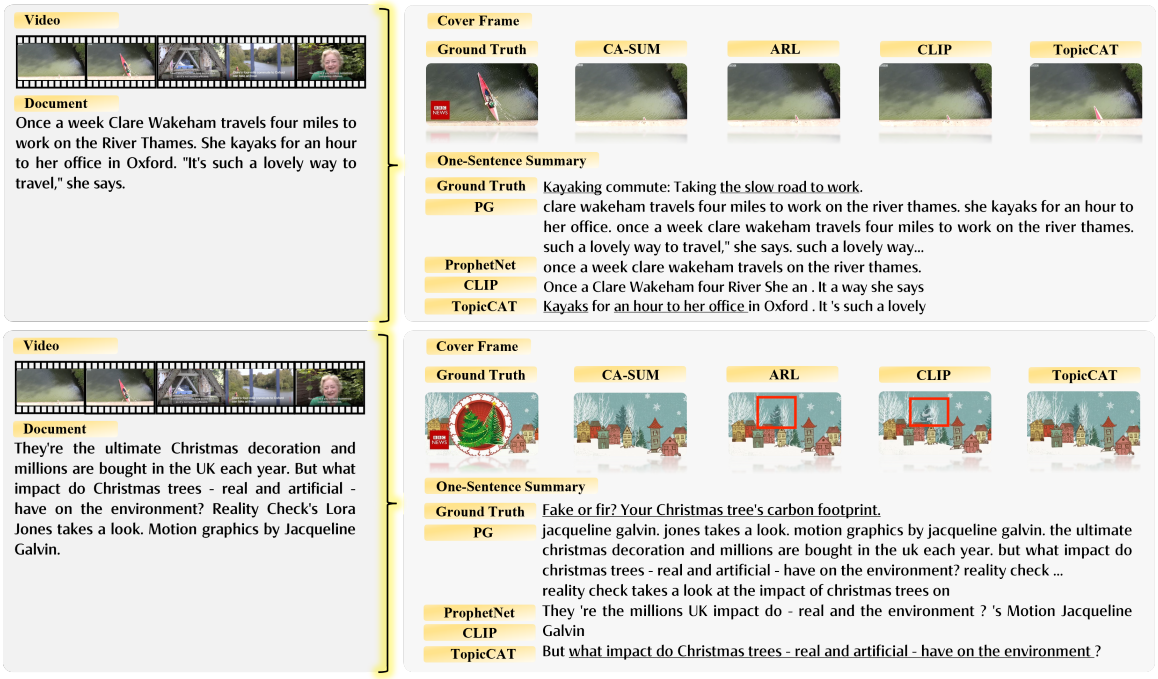

FIGURE 6.4. Example summaries generated by the baseline methods and TopicCAT.

Figure 5.5 compares the summaries produced by our TopicCAT and baselines, and a reference summary. The examples demonstrate that TopicCAT produces factually accurate summaries,

successfully identifying key information that aligns with the reference summary, as underlined. Remarkably, TopicCAT achieves this without the need for supervision. In contrast, the baselines fall short of capturing the essence of the documents. TopicCAT's textual summary is reasonably fluent yet has a room for further improvement.

In the kayaking example, TopicCAT aligns closely with the ground truth by recognizing the kayaking event on the river, a detail missed by CA-SUM, ARL, and CLIP. In the Christmas tree example, TopicCAT selects a quality frame showcasing the tree, whereas ARL and CLIP choose a watermark frame, as highlighted by a bounding box. Despite this success, there remains room for improvement in identifying key human-centered aspects like frame quality and central object emphasis.

In a broader research perspective of summarisation, while existing methods are more capable of advanced reasoning, they still remain constrained by the same limitations as their predecessors. Unfaithful or nonsensical output can be generated if the expected outcome is not clearly known or defined beforehand. This risk is present to some extent in almost every application that involves content generation. This risk is similar to human agents who can make mistakes and therefore need scripts and guidelines to assist them. As the current research direction seems to be moving away from the rule- and template-based approaches and towards more open discussions, an intriguing future direction would be to design a framework that achieves a balance between the two.

6.3.7 Visualisation of the Topic Space

We explore and visualise the latent topic space inferred by our multimodal topic model, mapping it into a 2D space using t-SNE. As shown in Figure 6.5, when coloured by modality, no single modality appears isolated in either the CLIP space or the multimodal topic space. Furthermore, in the visualisation distinguished by cluster, documents (blue) and videos (orange) tend to form more distinct clusters within the multimodal topic space generated by our model compared to the CLIP embedding space.

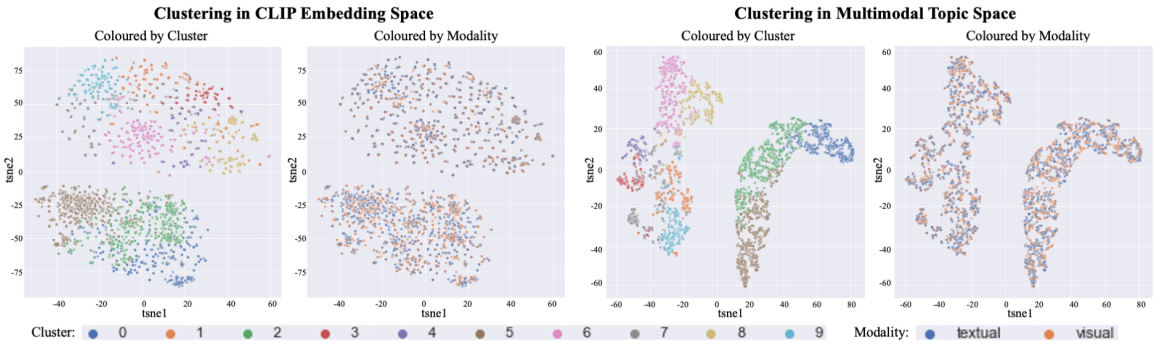

FIGURE 6.5. t-SNE visualization forthe topic distributions in the CLIP embedding space and those in our multimodal topic space.

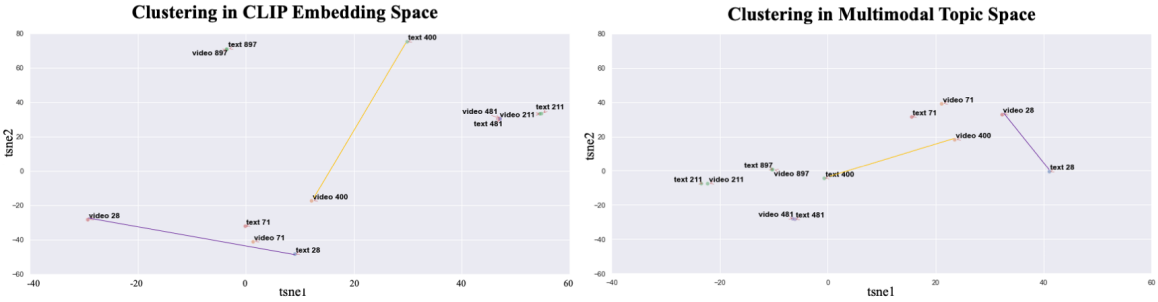

FIGURE 6.6. t-SNE visualizations of topic distributions of random video-document pairs.

In addition, we examine the proximity between documents and videos within the same pair as they are mapped into the latent topic space. A couple of randomly selected document-video pairs are illustrated in Figure 6.6. Generally, documents are found to be closer to their corresponding videos in themultimodal topic space than in the CLIP space, as indicated by the yellow and purple lines. These observations align with the intended behavior of our proposed mechanisms.

6.3.8 Limitations

Since the extreme multimodal summarisation task is a new field, prior datasets [52, 141] are not suitable for our modelling pipeline and evaluations. Our current evaluations rely on a single-source dataset from BBC News, possibly introducing style bias. Future research should expand to consider summarisation from diverse content, possibly through approaches like few-shot or zero-shot learning, leveraging large-scale pre-trained models. In addition,

our current optimal transport optimization utilizes a straightforward metric space for the cost function. Future work should investigate alternative approaches such as tree-sliced or graph-sliced metric spaces to enhance semantic distribution understanding and computational efficiency.

6.4 Conclusion

This chapter introduces TopicCAT, a novel deep learning method for extreme multimodal summarization, which condenses a video-document pair into a concise summary comprising a single cover frame and a sentence. TopicCAT is structured around two learning stages for an in-depth multimodal understanding guided by topics: a unimodal stage and a cross-modal stage, wherein a cross-modal topic model identifies the overarching themes within both documents and videos. Leveraging an optimal transport-based unsupervised learning, the approach optimizes summary coverage at the topic-level from a semantic distribution perspective. Extensive experiments on a large-scale extreme multimodal summarization dataset demonstrate the effectiveness of the proposed method.

CHAPTER 7

Conclusion

7.1 Summary and Conclusions

Text summarisation is a challenging task in natural language processing where the aim is to shorten a document into a brief and concise summary that captures the key information with as little redundancy as possible. This is helpful for people who want to quickly understand important information without reading long texts. As the amount of multimedia data increases, there is a growing interest in multimodal summarisation with multimodal output, which combines a video-document pair into a visual-textual summary. This type of summarisation provides users with a more comprehensive and visual understanding of events. While various methods have shown promising results, they have drawbacks such as high cost, low interpretability, and lack of conciseness. Therefore, this thesis addresses the gaps by devising unsupervised and interpretable text summarisation and multimodal summarisation methods.

Firstly, a non-learning-based extractive text summarisation method - OTExtSum is proposed by treating the text summarisation task as an optimal transport problem for the first time. Two optimisation strategies for OTExtSum are designed to optimise this problem formulation: beam search strategy and binary integer programming strategy.

Secondly, the first unsupervised compressive text summarisation method with dual-agent reinforcement learning, URLComSum, is proposed. An efficient and interpretable multi-head attentional pointer-based neural network is designed to learn the representation and extract salient sentences and words. The unsupervised reinforcement learning strategy is designed to

mimic human judgment by optimising summary quality in terms of the semantic coverage reward, measured by Wasserstein distance, and the fluency reward, measured by Syntactic Log-Odds Ratio (SLOR).

Thirdly, a new task - eXtreme Multimodal Summarisation with Multiple Output (XMSMO) is proposed. It aims to summarise a video-document pair into an extreme multimodal summary (i.e., one cover frame as the visual summary and one sentence as the textual summary). A novel unsupervised Hierarchical Optimal Transport Network (HOT-Net) is proposed. The hierarchical encoding and decoding are conducted across visual and textual modalities, and optimal transport solvers are introduced to guide the summaries to maximise their semantic coverage. A new large-scale dataset XMSMO-News is constructed for the research community to facilitate research in this new direction.

Finally, A novel transformer architecture - Topic-Guided Co-Attention Transformer (Topic-CAT) - is proposed for emerging extreme multimodal summarisation. It constructs a two-stage learning strategy for unimodal and cross-modal modelling with clustering-based cross-modal topic guidance. A novel optimal transport-guided unsupervised training strategy is devised to optimise TopicCAT from the perspective of the similarity between semantic distributions of topics.

7.2 Future Outlook

The potential research directions to advance text and multimodal summarisation are outlined as follows:

7.2.1 Explainable and faithful summarisation

Designing explainable and faithful summarisation methods becomes crucial as the demand for reliable and interpretable AI systems increases. These techniques are vital in building confidence in AI systems that aid decision-making and content curation [56] . By designing summarisation systems that are transparent in the generation process and accurate in terms

of the summary content, these methods can enable users to comprehend the reasoning and enhance the dependability of such systems.

7.2.2 Application-oriented and domain-specific summarisation

As summarisation methods have been advanced and are achieving promising results, these methods could be applied and customised to design application-oriented and domain-specific summarisation methods to tackle real-world needs. Examples are news summarisation [25], financial document summarisation [21], lifelogging summarisation [114], and text and multimodal simplification [112].

7.2.3 Query-base multimodal summarisation

There has been research on query-based text summarisation [90] and video summarisation [121] for decades. However, to my best knowledge, there is no existing research on query-based multimodal summarisation. Introducing user interaction in the summarisation process could better address users' information needs and help improve user satisfaction.

7.2.4 Evaluation of extreme multimodal summarisation

The current methods for automatically evaluating textual summaries focus on matching words and similarity against ground-truth or reference summaries and often overlook important qualities such as brevity, language quality, and factuality of a summary. Extreme summaries are brief, with an average of 23 tokens in the extreme text summarisation dataset XSum [80] and 12 tokens in this newly collected extreme multimodal summarisation dataset. Evaluating conciseness, language quality, and accuracy is crucial to gain deeper insights. However, evaluating these qualities is difficult because human evaluation is subjective, as pointed out in [36]. To overcome this challenge, creating automatic metrics for evaluating conciseness, language quality, and factuality would be crucial in future research to obtain more meaningful evaluation results.

Bibliography

[1] Jason Altschuler, Jonathan Weed and Philippe Rigollet. 'Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration'. In: *International Conference on Neural Information Processing Systems(NeurIPS)*. 2017.

[2] Evlampios Apostolidis et al. 'Combining Adversarial and Reinforcement Learning for Video Thumbnail Selection'. In: *International Conference on Multimedia Retrieval (ICMR)*. Taipei, Taiwan: Association for Computing Machinery, 2021, pp. 1–9.

[3] Evlampios Apostolidis et al. 'Summarizing Videos Using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames'. In: *International Conference on Multimedia Retrieval (ICMR)*. Newark, NJ, USA: Association for Computing Machinery, 2022, pp. 407–415.

[4] Evlampios Apostolidis et al. 'Video Summarization Using Deep Neural Networks: A Survey'. In: *Proceedings of the IEEE* (2021).

[5] W Ashworth. 'Abstracting as a fine art'. In: *Information scientist* (1973).

[6] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 'Neural machine translation by jointly learning to align and translate'. In: (2015).

[7] Isabel Cachola et al. 'TLDR: Extreme Summarization of Scientific Documents'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 4766–4777.

[8] Jiacheng Chen et al. 'Learning the best pooling strategy for visual semantic embedding'. In: *IEEE conference on computer vision and pattern recognition (CVPR)*. 2021.

[9] Jingqiang Chen and Hai Zhuge. 'Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN'. In: *Conference on Empirical Methods in*

Natural Language Processing (EMNLP). Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4046–4056.

[10] Liqun Chen et al. 'Adversarial Text Generation via Feature-Mover's Distance'. In: *International Conference on Neural Information Processing Systems (NeurIPS)*. 2018.

[11] Yen-Chun Chen and Mohit Bansal. 'Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 675–686.

[12] Arman Cohan et al. 'A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2018.

[13] Shrey Desai, Jiacheng Xu and Greg Durrett. 'Compressive Summarization with Plausibility and Salience Modeling'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[14] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2019.

[15] Yue Dong et al. 'BanditSum: Extractive Summarization as a Contextual Bandit'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2018.

[16] H. P. Edmundson. 'New Methods in Automatic Extracting'. In: *Journal of the ACM* 16 (1969), pp. 264–285.

[17] G. Erkan and D. R. Radev. 'LexRank: Graph-based Lexical Centrality as Salience in Text Summarization'. In: *Journal of Artificial Intelligence Research* (2004).

[18] Alexander Fabbri et al. 'Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

[19] Jean Feydy et al. ‘Interpolating between Optimal Transport and MMD using Sinkhorn Divergences’. In: *International Conference on Artificial Intelligence and Statistics*. 2019.

[20] Katja Filippova. ‘Multi-Sentence Compression: Finding Shortest Paths in Word Graphs’. In: *International Conference on Computational Linguistics (COLING)*. 2010.

[21] Negar Foroutan et al. ‘Multilingual Text Summarization on Financial Documents’. In: *Financial Narrative Processing Workshop*. Marseille, France: European Language Resources Association, June 2022, pp. 53–58.

[22] Mahak Gambhir and Vishal Gupta. ‘Recent automatic text summarization techniques: a survey’. In: *Artificial Intelligence Review* (2017).

[23] Kavita Ganesan, ChengXiang Zhai and Jiawei Han. ‘Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions’. In: *International Conference on Computational Linguistics (COLING)*. 2010.

[24] Shima Gerani et al. ‘Abstractive Summarization of Product Reviews Using Discourse Structure’. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[25] Demian Gholipour Ghalandari and Georgiana Ifrim. ‘Examining the State-of-the-Art in News Timeline Summarization’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, July 2020, pp. 1322–1334.

[26] Dan Gillick and Benoit Favre. ‘A scalable global model for summarization’. In: *Workshop on Integer Linear Programming for Natural Language Processing*. 2009.

[27] Yihong Gong and Xin Liu. ‘Generic text summarization using relevance measure and latent semantic analysis’. In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 2001.

[28] Alex Graves and Jürgen Schmidhuber. ‘Framewise phoneme classification with bidirectional LSTM and other neural network architectures’. In: *Neural networks* (2005).

[29] Max Grusky, Mor Naaman and Yoav Artzi. 'Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2018.

[30] Hongxiang Gu and Viswanathan Swaminathan. 'From Thumbnails to Summaries-A Single Deep Neural Network to Rule Them All'. In: *IEEE International Conference on Multimedia and Expo (ICME)*. San Diego, CA, USA: IEEE, 2018, pp. 1–6.

[31] Genliang Guan et al. 'Keypoint-based keyframe selection'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.4 (2013), pp. 729–734.

[32] Vishal Gupta and Gurpreet Singh Lehal. 'A survey of text summarization extractive techniques'. In: *Journal of emerging technologies in web intelligence* (2010).

[33] Sanda Harabagiu, Finley Lacatusu and Paul Morarescu. 'Multidocument Summarization with GISTexter'. In: *International Conference on Language Resources and Evaluation (LREC)*. 2002.

[34] Karl Moritz Hermann et al. 'Teaching Machines to Read and Comprehend'. In: *International Conference on Neural Information Processing Systems (NeurIPS)*. Montreal, Canada: MIT Press, 2015, pp. 1693–1701.

[35] Eduard Hovy and Chin-Yew Lin. 'Automated Text Summarization and the Summarist System'. In: *TIPSTER Text Program Phase III*. 1998.

[36] David M. Howcroft et al. 'Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions'. In: *International Conference on Natural Language Generation (INLG)*. Dublin, Ireland: Association for Computational Linguistics, 2020, pp. 169–182.

[37] Cheng Huang and Hongmei Wang. 'A Novel Key-Frames Selection Framework for Comprehensive Video Summarization'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.2 (2020), pp. 577–589.

[38] Anubhav Jangra et al. 'A Survey on Multi-Modal Summarization'. In: *ACM Computing Surveys* (Feb. 2023). ISSN: 0360-0300.

[39] Zhong Ji et al. ‘Video summarization with attention-based encoder–decoder networks’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.6 (2019), pp. 1709–1717.

[40] Zhong Ji et al. ‘Video Summarization With Attention-Based Encoder–Decoder Networks’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.6 (2020), pp. 1709–1717.

[41] Yifan Jiao et al. ‘Three-Dimensional Attention-Based Deep Ranking Model for Video Highlight Detection’. In: *IEEE Transactions on Multimedia* 20.10 (2018), pp. 2693–2705.

[42] Katharina Kann, Sascha Rothe and Katja Filippova. ‘Sentence-Level Fluency Evaluation: References Help, But Can Be Spared!’ In: *Conference on Computational Natural Language Learning (CoNLL)*. 2018.

[43] Wafaa S. El-Kassas et al. ‘Automatic text summarization: A comprehensive survey’. In: *Expert Systems with Applications* (2021).

[44] Anastassia Kornilova and Vladimir Eidelman. ‘BillSum: A Corpus for Automatic Summarization of US Legislation’. In: *Workshop on New Frontiers in Summarization (NFiS)*. 2019.

[45] Krtin Kumar and Jackie Chi Kit Cheung. ‘Understanding the Behaviour of Neural Abstractive Summarizers using Contrastive Examples’. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2019.

[46] Julian Kupiec, Jan Pedersen and Francine Chen. ‘A trainable document summarizer’. In: *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 1995.

[47] Matt Kusner et al. ‘From Word Embeddings To Document Distances’. In: *International Conference on Machine Learning (ICML)*. Lille, France: JMLR.org, 2015, pp. 957–966.

[48] Philippe Laban et al. 'The Summary Loop: Learning to Write Abstractive Summaries Without Examples'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, 2020, pp. 5135–5150.

[49] Jey Han Lau, Alexander Clark and Shalom Lappin. 'Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge'. In: *Cognitive Science* 41.5 (2017), pp. 1202–1241.

[50] Mike Lewis et al. 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020.

[51] Haoran Li et al. 'Multi-Modal Sentence Summarization with Modality Attention and Image Filtering'. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI'18. Stockholm, Sweden: AAAI Press, 2018, pp. 4152–4158. ISBN: 9780999241127.

[52] Mingzhe Li et al. 'VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 9360–9369.

[53] Kexin Liao, Logan Lebanoff and Fei Liu. 'Abstract Meaning Representation for Multi-Document Summarization'. In: *International Conference on Computational Linguistics (COLING)*. 2018.

[54] Chin-Yew Lin. 'ROUGE: A Package for Automatic Evaluation of Summaries'. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[55] Chin-Yew Lin and Eduard Hovy. 'The Automated Acquisition of Topic Signatures for Text Summarization'. In: *International Conference on Computational Linguistics (COLING)*. 2000.

[56] Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. 'Explainable AI: A Review of Machine Learning Interpretability Methods'. In: *Entropy* 23.1 (2021). ISSN: 1099-4300.

[57] Fei Liu et al. ‘Toward Abstractive Summarization Using Semantic Representations’. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2015.

[58] Liyuan Liu et al. ‘Understanding the Difficulty of Training Transformers’. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[59] Nayu Liu et al. ‘Multistage Fusion with Forget Gate for Multimodal Summarization in Open-Domain Videos’. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1834–1845.

[60] Yang Liu and Mirella Lapata. ‘Text Summarization with Pretrained Encoders’. In: *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3730–3740.

[61] Yixin Liu et al. ‘BRIO: Bringing Order to Abstractive Summarization’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2890–2903.

[62] Elena Lloret and Manuel Palomar. ‘Text summarisation in progress: a literature review’. In: *Artificial Intelligence Review* (2012).

[63] Ilya Loshchilov and Frank Hutter. ‘Decoupled Weight Decay Regularization’. In: *International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA: OpenReview.net, 2018.

[64] H. P. Luhn. ‘A Business Intelligence System’. In: *IBM Journal of Research and Development* (1958).

[65] Jiebo Luo, Christophe Papin and Kathleen Costello. ‘Towards extracting semantically meaningful key frames from personal video clips: from humans to computers’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 19.2 (2008), pp. 289–301.

[66] Ling Luo et al. ‘Reading Like HER: Human Reading Inspired Extractive Summarization’. In: *Conference on Empirical Methods in Natural Language Processing and*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[67] Caixia Ma et al. 'Adaptive Multiview Graph Difference Analysis for Video Summarization'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.12 (2022), pp. 8795–8808.

[68] Mingyang Ma et al. 'Similarity based block sparse subset selection for video summarization'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.10 (2020), pp. 3967–3980.

[69] Chris J. Maddison, Andriy Mnih and Yee Whye Teh. 'The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables'. In: *International Conference on Learning Representations (ICLR)*. 2016.

[70] Inderjeet Mani. *Automatic summarization*. John Benjamins Publishing, 2001.

[71] Daniel Marcu. 'Discourse Trees Are Good Indicators of Importance in Text'. In: *Advances in Automatic Text Summarization*. 1999.

[72] Joshua Maynez et al. 'On Faithfulness and Factuality in Abstractive Summarization'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Online: Association for Computational Linguistics, 2020, pp. 1906–1919.

[73] Ryan McDonald. 'A Study of Global Inference Algorithms in Multi-Document Summarization'. In: *European Conference on IR Research*. 2007.

[74] Clara Meister, Ryan Cotterell and Tim Vieira. 'If Beam Search Is the Answer, What Was the Question?' In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[75] Afonso Mendes et al. 'Jointly Extracting and Compressing Documents with Summary State Representations'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2019.

[76] Safa Messaoud et al. 'DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization'. In: *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1389–1399.

[77] Rada Mihalcea and Paul Tarau. 'TextRank: Bringing Order into Text'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 404–411.

[78] Ibrahim F. Moawad and Mostafa M. Aref. 'Semantic graph reduction approach for abstractive Text Summarization'. In: *International Conference on Computer Engineering and Systems (ICCES)* (2012).

[79] Ramesh Nallapati, Feifei Zhai and Bowen Zhou. 'Summarunner: A recurrent neural network based sequence model for extractive summarization of documents'. In: *AAAI Conference on Artificial Intelligence*. 2017.

[80] Shashi Narayan, Shay B. Cohen and Mirella Lapata. 'Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1797–1807.

[81] Shashi Narayan, Shay B. Cohen and Mirella Lapata. 'Ranking Sentences for Extractive Summarization with Reinforcement Learning'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2018.

[82] Shashi Narayan et al. 'Stepwise Extractive Summarization and Planning with Structured Transformers'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[83] Ani Nenkova and Kathleen McKeown. 'Automatic Summarization'. In: *Foundations and Trends® in Information Retrieval* 5 (2011), pp. 103–233.

[84] Andrew Y Ng. 'Feature selection, L 1 vs. L 2 regularization, and rotational invariance'. In: *International Conference on Machine Learning (ICML)*. 2004.

[85] Chikashi Nobata and Satoshi Sekine. 'CRL/NYU summarization system'. In: *Document Understanding Conference (DUC)*. 2004.

[86] Tatsuro Oya et al. 'A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships'. In: *International Natural Language Generation Conference (INLG)*. 2014.

[87] Vishakh Padmakumar and He He. 'Unsupervised Extractive Summarization using Pointwise Mutual Information'. In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2021.

[88] Shruti Palaskar et al. 'Multimodal Abstractive Summarization for How2 Videos'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 6587–6596.

[89] Jacob Parnell, Inigo Jauregi Unanue and Massimo Piccardi. 'RewardsOfSum: Exploring Reinforcement Learning Rewards for Summarisation'. In: *Workshop on Structured Prediction for NLP (SPNLP)*. Online: Association for Computational Linguistics, 2021, pp. 1–11.

[90] Ramakanth Pasunuru et al. 'Data Augmentation for Abstractive Query-Focused Multi-Document Summarization'. In: *AAAI Conference on Artificial Intelligence*. Vol. 35. 15. May 2021, pp. 13666–13674.

[91] Adam Pauls and Dan Klein. 'Large-Scale Syntactic Language Modeling with Treelets'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 959–968.

[92] Romain Paulus, Caiming Xiong and Richard Socher. 'A Deep Reinforced Model for Abstractive Summarization'. In: *International Conference on Learning Representations (ICLR)*. 2018.

[93] Jeffrey Pennington, Richard Socher and Christopher Manning. 'Glove: Global Vectors for Word Representation'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[94] Gabriel Peyré, Marco Cuturi et al. 'Computational optimal transport: With applications to data science'. In: *Foundations and Trends in Machine Learning* (2019).

[95] Laura Plaza, Alberto Díaz and Pablo Gervás. 'A semantic graph-based approach to biomedical summarisation'. In: *Artificial Intelligence in Medicine* (2011).

[96] Martin Popel and Ondřej Bojar. 'Training Tips for the Transformer Model'. In: *The Prague Bulletin of Mathematical Linguistics (NeurIPS)* (2018).

[97] Horst Pöttker. 'News and its communicative quality: the inverted pyramid—when and why did it appear?' In: *Journalism Studies* 4.4 (2003), pp. 501–511.

[98] Kyle Pretorious and Nelishia Pillay. 'A Comparative Study of Classifiers for Thumbnail Selection'. In: *International Joint Conference on Neural Networks (IJCNN)*. Glasgow, United Kingdom, IEEE, 2020, pp. 1–7.

[99] Weizhen Qi et al. 'ProphetNet: Predicting Future N-gram for Sequence-to-SequencePre-training'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020, pp. 2401–2410.

[100] Dragomir R. Radev, Hongyan Jing and Malgorzata Budzikowska. 'Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies'. In: *Conference of the North American Chapter of the Association for Computational Linguistics - Applied Natural Language Processing Conference Workshop: Automatic Summarization (NAACL-ANLP-AutoSum)*. 2000.

[101] Alec Radford et al. 'Learning Transferable Visual Models From Natural Language Supervision'. In: *International Conference on Machine Learning (ICML)*. Ed. by Marina Meila and Tong Zhang. Online: PMLR, 2021, pp. 8748–8763.

[102] Jian Ren et al. 'Best Frame Selection in a Short Video'. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Snowmass, CO, USA: IEEE, 2020, pp. 3201–3210.

[103] Steven J Rennie et al. 'Self-critical sequence training for image captioning'. In: *IEEE conference on computer vision and pattern recognition (CVPR)*. 2017.

[104] Alexander M. Rush, Sumit Chopra and Jason Weston. 'A Neural Attention Model for Abstractive Sentence Summarization'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2015.

[105] Horacio Saggion. 'Automatic summarization: an overview'. In: *Revue française de linguistique appliquée* (2008).

[106] Abigail See, Peter J. Liu and Christopher D. Manning. 'Get To The Point: Summarization with Pointer-Generator Networks'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1073–1083.

[107] Aidean Sharghi, Jacob S Laurel and Boqing Gong. 'Query-focused video summarization: Dataset, evaluation, and a memory network based approach'. In: *IEEE conference*

on computer vision and pattern recognition (CVPR). Honolulu, HI, USA: IEEE, 2017, pp. 2127–2136.

[108] Beaux Sharifi, Mark-Anthony Hutton and Jugal Kalita. 'Summarizing Microblogs Automatically'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2010.

[109] Karen Spärck Jones. 'Automatic summarising: The state of the art'. In: *Information Processing and Management* (2007).

[110] Ilya Sutskever et al. 'On the importance of initialization and momentum in deep learning'. In: *International Conference on Machine Learning (ICML)*. 2013.

[111] Kyle Swanson, Lili Yu and Tao Lei. 'Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020.

[112] Suha S. Al-Thanyyan and Aqil M. Azmi. 'Automated Text Simplification: A Survey'. In: *ACM Computing Survey* 54.2 (Mar. 2021). ISSN: 0360-0300.

[113] Christoph Tillmann and Hermann Ney. 'Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation'. In: *Computational Linguistics* (2003).

[114] V. Javier Traver and Dima Damen. 'Egocentric video summarisation via purpose-oriented frame scoring and selection'. In: *Expert Systems with Applications* 189 (2022), p. 116079. ISSN: 0957-4174.

[115] Ashish Vaswani et al. 'Attention is all you need'. In: *International Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 30. Long Beach, CA, USA: Curran Associates, Inc., 2017.

[116] Petar Veličković et al. 'Graph Attention Networks'. In: *International Conference on Learning Representations (ICLR)*. 2018.

[117] Oriol Vinyals, Meire Fortunato and Navdeep Jaitly. 'Pointer networks'. In: *International Conference on Neural Information Processing Systems (NeurIPS)*. 2015.

[118] Junbo Wang et al. 'Stacked memory network for video summarization'. In: *ACM International Conference on Multimedia (ACMMM)*. 2019.

[119] Lihan Wang et al. ‘Abstractive Text Summarization with Hierarchical Multi-Scale Abstraction Modeling and Dynamic Memory’. In: *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 2086–2090.

[120] Thomas Wolf et al. ‘Transformers: State-of-the-Art Natural Language Processing’. In: *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[121] Guande Wu, Jianzhe Lin and Claudio T. Silva. ‘IntentVizor: Towards Generic Query Guided Interactive Video Summarization’. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10503–10512.

[122] Jiacheng Xu and Greg Durrett. ‘Neural Extractive Text Summarization with Syntactic Compression’. In: *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

[123] Jiacheng Xu et al. ‘Discourse-Aware Neural Extractive Text Summarization’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020.

[124] Jingjing Xu et al. ‘Vocabulary Learning via Optimal Transport for Neural Machine Translation’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2021.

[125] Jin-ge Yao, Xiaojun Wan and Jianguo Xiao. ‘Recent advances in document summarization’. In: *Knowledge and Information Systems* 53.2 (2017), pp. 297–336.

[126] Mehdi Yousfi-Monod and Violaine Prince. ‘Sentence Compression as a Step in Summarization or an Alternative Path in Text Shortening’. In: *International Conference on Computational Linguistics (COLING)*. 2008.

[127] Li Yuan et al. ‘Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks’. In: *IEEE Transactions on Multimedia* 22.10 (2020), pp. 2711–2722.

[128] Ye Yuan and Jiawan Zhang. ‘Unsupervised Video Summarization via Deep Reinforcement Learning With Shot-Level Semantics’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.1 (2023), pp. 445–456.

[129] Ye Yuan and Jiawan Zhang. 'Unsupervised Video Summarization via Deep Reinforcement Learning with Shot-level Semantics'. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[130] Mikhail Yurochkin et al. 'Hierarchical optimal transport for document representation'. In: *International Conference on Neural Information Processing Systems (NeurIPS)*. 2019.

[131] Jingqing Zhang et al. 'Pegasus: Pre-training with extracted gap-sentences for abstractive summarization'. In: *International Conference on Machine Learning (ICML)*. Online: PMLR, 2020, pp. 11328–11339.

[132] Litian Zhang et al. 'Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization'. In: *AAAI Conference on Artificial Intelligence*. Newark, NJ, USA: Association for Computing Machinery, 2022, pp. 239–248.

[133] Tianyi Zhang et al. 'BERTScore: Evaluating Text Generation with BERT'. In: *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia: OpenReview.net, 2020.

[134] Xingxing Zhang, Furu Wei and Ming Zhou. 'HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization'. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

[135] Xingxing Zhang et al. 'Neural Latent Extractive Document Summarization'. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2018.

[136] Zihan Zhang et al. 'Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics'. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Seattle, United States: Association for Computational Linguistics, 2022, pp. 3886–3893.

[137] Bin Zhao, Xuelong Li and Xiaoqiang Lu. 'Hierarchical Recurrent Neural Network for Video Summarization'. In: *ACM International Conference on Multimedia*. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 863–871. ISBN: 9781450349062.

[138] Hao Zheng and Mirella Lapata. ‘Sentence Centrality Revisited for Unsupervised Summarization’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

[139] Ming Zhong et al. ‘Extractive Summarization as Text Matching’. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020.

[140] Huiyu Zhou et al. ‘Feature extraction and clustering for dynamic video summarisation’. In: *Neurocomputing* 73.10 (2010), pp. 1718–1729. ISSN: 0925-2312.

[141] Junnan Zhu et al. ‘MSMO: Multimodal Summarization with Multimodal Output’. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4154–4164.

[142] Junnan Zhu et al. ‘Multimodal summarization with guidance of multimodal reference’. In: *AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 9749–9756.

[143] Wenwu Zhu, Xin Wang and Hongzhi Li. ‘Multi-modal deep analysis for multimedia’. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.10 (2019), pp. 3740–3764.