

**FACULTY
OF MATHEMATICS
AND PHYSICS**
Charles University

DOCTORAL THESIS

Mateusz Krubiński

Multimodal Summarization

Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina, Ph.D.

Study Program: Computational Linguistics

Prague 2024

I declare that I carried out this doctoral thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

Prague, July 15, 2024

Mateusz Krubiński

Acknowledgments

Mojej żonie, Marii, bez której ta przygoda by się nie rozpoczęła ani, tym bardziej, nie zakończyła.

I want to thank my supervisor, doc. RNDr. Pavel Pecina, Ph.D., for all the help and support and for the “artistic freedom” that he entrusted me with.

Throughout the duration of my doctoral studies, my research was supported by: i) the European Commission via its H2020 Program (contract no. 870930); ii) CELSA (project no. 19/018); iii) the Czech Science Foundation (grant no. 19-26934X); iv) the Charles University (GAUK 291923 and SVV 260 698), and has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (projects no. LM2018101 and no. LM2023062).

Title: Multimodal Summarization

Author: Mateusz Krubiński

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Pavel Pecina, Ph.D.,
Institute of Formal and Applied Linguistics

Abstract: The task of Multimodal Summarization aims to fuse disjoint information from several sources (modalities) and distill it into a concise and precise summary. In our research, we approach a text-centric variant that requires textual content in both the input and the output, i.e., the summary. Specifically, we focus on the Multimodal Summarization with Multimodal Output (MSMO) approach, which summarizes a textual document accompanied by either a collection of images or a short video into a textual summary accompanied by a single image. On the modeling side, we are interested in supervised formulations that explore a single neural model to generate the multimodal summary end-to-end, i.e., by simultaneously processing textual and visual modalities. Considering the task’s novelty, it still lacks the core components of a well-established field, such as standardized benchmarks (datasets), publicly available baseline models, and even task-specific metrics. Therefore, our main contributions are aimed at performing basic research to establish foundations for future work. Namely, we: i) curate and publish a large-scale video-based dataset for MSMO; ii) perform experiments to establish the role of pre-training and the influence of the (quality of) visual input on the (quality of) textual output; iii) design a human evaluation framework for MSMO, and propose a novel metric for evaluating the quality of textual output; iv) propose a simplified, multi-task formulation of MSMO, that unifies the image-based, video-based, and text-only variants with a single architecture.

Keywords: summarization, text summarization, vision-language modeling, multimodal data

Contents

Introduction	3
1 Background	5
1.1 Text Summarization	6
1.1.1 Abstractive Summarization	6
1.1.2 Extractive Summarization	8
1.2 Vision and Language Modeling	10
1.2.1 Feature encoding	10
1.2.2 Feature fusion	12
2 Multimodal Summarization	15
2.1 Multimodal Summarization with Unimodal Output	18
2.1.1 Multiple documents \rightarrow Text	18
2.1.2 Text + Image(s) \rightarrow Text	18
2.1.3 Text + Video \rightarrow Text	21
2.1.4 Other formulations	23
2.2 Multimodal Summarization with Multimodal Output	24
2.2.1 Text + Images \rightarrow Text + Image	24
2.2.2 Text + Video \rightarrow Text + Image	26
2.2.3 Other formulations	28
3 Quality Evaluation	29
3.1 Textual Output	30
3.1.1 Human Evaluation	31
3.1.2 String-based metrics	35
3.1.3 Embedding-based metrics	36
3.1.4 QA-based metrics	37
3.1.5 LLM-based metrics	38
3.1.6 Trainable metrics	39
3.1.7 COMES	41
3.2 Visual Output	46
3.3 Multimodal Output	48
4 Datasets	53
4.1 Overview	53
4.2 MLASK	55
4.3 Extension of the M3LS dataset	59
5 Experiments	63
5.1 MLASK – MMS	63
5.1.1 Motivation and Overview	63
5.1.2 Implementation	66
5.1.3 Results and Ablation Studies	69
5.1.4 Implications	73
5.2 Unifying Uni- and Multi-modal Summarization	75
5.2.1 Motivation and Overview	75

5.2.2	Implementation	77
5.2.3	Results	81
5.2.4	Ablation Studies	84
Conclusions		89
Bibliography		93
Appendix A MTEQA		121
A.1	Overview	121
A.2	Implementation	122
A.3	Discussion	125
Appendix B Auxiliary Results		129
B.1	COMES	129
B.2	Examples of Model Outputs	132
B.2.1	MLASK-MMS	132
B.2.2	UNMGH	135
List of Abbreviations		139
List of Figures		141
List of Tables		145
List of Publications		147

Introduction

As per Oxford English Dictionary¹:

modality (pl. *modalities*) – the particular way in which something exists, is experienced or is done

summary (pl. *summaries*) – a short statement that gives only the main points of something, not the details

The task of automatic summarization aims to reduce the *size* of input data in a way that preserves the *key information* via the usage of an automated, *computational process*. The exact meaning of *size* or the notion of *key information* or *computational process* will heavily depend on the particular kind of data/problem that we are currently dealing with. When summarizing a piece of longer text (e.g., a news article or a medical report), it is often the case that the summary is also a piece of text in a similar format, just shorter. Therefore, the *size* can be expressed in terms of words or characters. But what about summarizing, e.g., logs from an internet service? In practical applications, a useful summary should probably consist of some aggregated statistics (e.g., the average number of API calls per minute, a distribution with regards to the age/gender/occupation of users that are interacting with the service), but also explicitly list events that must be manually reviewed (e.g., a time-window with a high proportion of failed HTTP requests). In that case, the exact form of input/summary will be different – how to define *size*? Looking from a different perspective – how to define *key information* when summarizing a piece of poetry, a poem? Approaching the problem from yet another angle – when summarizing a video, do we consider, e.g., downloading the video from the internet and transforming it into a standardized format (e.g., **mp4**) a part of *computational process*, or do we assume that such *process* starts with a sequence of frames (sampled from the video) which are ready to be consumed by a Machine Learning model?

Those kinds of questions become even more challenging once we take into account that the world around us – and the information that we consume – is *multimodal*. The notion of *modality* is not well defined and depends upon the particular context. One could say that an apple is *multimodal* since it has a texture (perceived with touch), a color (perceived with sight), a scent (perceived with smell), and a flavor (perceived with taste). Humans are not able to smell a color or identify a taste by touch. Therefore, distinguishing modalities by the particular sense that they stimulate introduces a categorization with well-defined categories and clear rules for separation. In that regard, a news article that one reads on their smartphone while eating breakfast is not multimodal – all the information is perceived with one’s eyes (sight). However, a news article will often consist of a mixture of text (paragraphs, image captions), images (photographs), and video (a clip with the recording of an event) – and those differ greatly from the technical perspective (e.g., how the data is stored, how it is embedded into a website, whether it can be transformed by screen readers for visually impaired, etc.). Those differences play a key role when we consider the task of summarizing such a *multimodal* news article (e.g., to present it on a newsfeed or to draw the

¹<https://www.oxfordlearnersdictionaries.com>

attention of the reader).

In this thesis, we approach the problem of automatic Multimodal Summarization, which aims to summarize information from multiple modalities into a common, concise summary. The modalities that we consider are text (that we require both in the input and in the summary) and vision (images and videos). We are interested in approaches (models) realized with neural networks and end-to-end formulations, i.e., ones that simultaneously process text and vision. From a scientific perspective, our research is motivated by the complexity of Vision and Language problems. While machine learning models are able to classify an image as either a cat or a dog with greater accuracy than humans or to generate translations that (human) annotators consider superior to human-generated ones, tasks that require reasoning over the combination of textual and visual modalities are still far from solving.

The thesis is structured as follows: in Chapter 1, we introduce the relevant background concerning Text Summarization and ViL modeling. In Chapter 2, we formally introduce the task of Multimodal Summarization and describe the problems and formulations approached in previous research. In Chapter 3, we look at the quality evaluation, i.e., what metrics are explored to estimate the quality of automatic summaries and what protocols are employed to collect human judgments regarding the quality of the automatically generated outputs. In Chapter 4, we look at the characteristics of publicly available datasets that enabled the previous research. Finally, in Chapter 5, we describe our experiments related to Multimodal Summarization. Our findings covered in this thesis are based on five conference papers that the author published during their doctoral studies (see “List of Publications” at the end of the thesis). A link between the thesis and published work is done via visual indicators, i.e.,

This section is based on the XYZ article.

to match a section or a chapter to a corresponding article (publication).

Specifically, the contributions presented in this thesis are as follows:

- a proposal of the COMES metric (see Section 3.1.7) for evaluating the quality of textual summaries;
- a novel framework for collecting human annotations judging the quality and relevance of pictorial summaries (see Section 3.3);
- a curation and a publication of a large-scale dataset for video-based Multimodal Summarization (see Section 4.2);
- an extension of an existing large-scale dataset for image-based Multimodal Summarization by enriching it with pictorial summaries (see Section 4.3);
- experiments aiming at establishing the role of task-specific pre-training and the influence of the visual input on the quality of textual output (see Section 5.1);
- a unified formulation of MSMO, merging text-only, image-based, and video-based problems with a common, encoder-decoder architecture trainable in a multi-task fashion (see Section 5.2);
- a novel metric dedicated to evaluating Machine Translation outputs (see Appendix A), inspired by the QA/QG approaches to textual summary evaluation.

1. Background

In this Chapter, we wish to establish a set of core concepts, ensuring that both we and the reader have a shared understanding of them.

In Section 1.1, we will consider Text Summarization (a core problem for text-centric Multimodal Summarization), highlighting the two variants explored in previous research. Namely, in Section 1.1.1, we will take a closer look at the abstractive (generative) approaches, and in Section 1.1.2, we will briefly cover the extractive ones. One might argue that Video Summarization is also a core problem, especially within the video-based formulation of MSMO (see Section 2.2). However, since the typical benchmark datasets (see, e.g., Apostolidis et al. (2021)) are annotated with (shot/scene-level) human preferences, and the target is formulated as a short clip (video skim), none of the methods are directly relevant/applicable to our research.

In Section 1.2, we will touch upon the broad family of ViL tasks that the Multimodal Summarization belongs to. Our concern will be with the methods commonly explored to obtain numerical representations of visual input (see Section 1.2.1) and the modeling approaches to combining textual and visual representations (see Section 1.2.2).

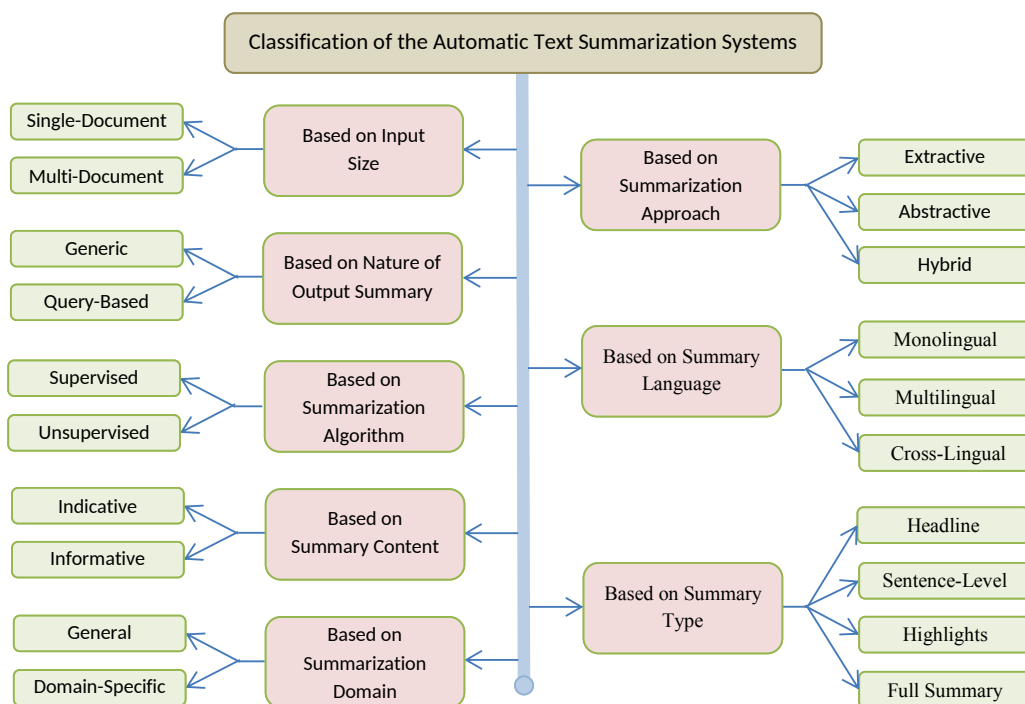


Figure 1.1: Categorization of automatic Text Summarization systems proposed by El-Kassas et al. (2021). Figure reprint from El-Kassas et al. (2021).

1.1 Text Summarization

The task of Text Summarization is one of the classical problems in Natural Language Processing, aiming at creating a concise and coherent *summary* of a textual input that preserves the key concepts and retains the essential pieces of information. Considering both its *depth* – early works published already in the 1950s and 1960s, e.g., Luhn (1958); Edmundson (1969) – and its *breadth* (see Figure 1.1), trying to provide even a concise overview, which would still do justice to every variety and sub-problem, is out of scope of this work. Instead, we will focus on the modeling approaches, limiting ourselves to those that are realized with neural networks. Specifically, we would like to provide a (subjective) overview of what we consider major paradigm shifts and how they interleave with breakthroughs in multimodal approaches (see Chapter 2). Additionally, we will limit our recap to single-document problems and approaches operating in a supervised manner, i.e., with a reference summary. For a comprehensive review of the classical automatic Text Summarization methods, we refer the reader to one of the following surveys: Suleiman and Awajan (2020); El-Kassas et al. (2021); Widyassari et al. (2022); Yadav et al. (2022); Zhang et al. (2022b). Readers interested in multi-document summarization may consult e.g., Ma et al. (2022), or newer works based on LLMs, e.g., Bhaskar et al. (2023) or Huang et al. (2023).

For the remainder of this Section, the input document D , the reference summary S , and the automatically created summary Y will be represented by a series of tokens (words), i.e., $D = (d_1, \dots, d_k)$; $S = (s_1, \dots, s_m)$; $Y = (y_1, \dots, y_n)$. The summarization model will refer to a system capable of indirectly (or directly) modeling the conditional probability of $p(Y|D)$. Unless stated otherwise, the decision process is to pick the summary Y from the pool of available candidates $\{\hat{Y}\}$ that maximizes the conditional probability, i.e., $Y = \arg \max_{\{\hat{Y}\}} p(\hat{Y}|D)$

1.1.1 Abstractive Summarization

Within the abstractive formulation of Text Summarization, the summary Y is generated – or built – based on the available components (tokens), conditioned on the input document D . The conditioning is realized with a recurrent neural network involving a decoder, i.e., a component capable of modeling the conditional, per-token probability. In that case, the pool of available candidates $\{\hat{Y}\}$ is not realized directly but estimated with a point-wise (token-based) modeling, i.e., $y_i = \arg \max_y p(y|y_1, \dots, y_{i-1}, D)$, based on the input document D and the previously generated tokens y_1, \dots, y_{i-1} . Such generation process is called autoregressive generation, as the probability of a consecutive token depends upon the tokens generated so far. The non-autoregressive generation that conditions only on the input document D (and on the positional index i) belongs to a separate sub-field (see, e.g., Su et al. (2021)) that we will not consider here. The generation process (decoding) does not assume the length of the output (summary) *a priori* but selects the consecutive tokens until the model itself decides to end the generation, as indicated by a special [EOS] token – or until the maximum desired length of summary is reached.

On one hand, the advances in the field follow the generic trends of sequence-to-sequence modeling. Namely, the replacement of the simple recurrent RNN cells with LSTM cells (Hochreiter and Schmidhuber, 1997), followed by the simpler but similarly performing GRU (Chung et al., 2014) cells. The introduction of the Transformer (Vaswani et al., 2017) architecture¹ enabled a direct ($O(1)$) interaction between tokens (compared to $O(n)$ for RNNs) at the cost of quadratic complexity ($O(n^2)$ for Transformer vs $O(n)$ for RNNs) of computing the forward pass. Devlin et al. (2019) proposed the masked language modeling to pre-train the textual encoder, contributing to the novel training pipelines – instead of training the models from scratch, it became a custom to fine-tune pre-trained components. The unified text-to-text formulation of pre-training – Raffel et al. (2020) proposed a variant based on the premises of transfer learning, while Lewis et al. (2020a) explored the denoising (autoencoder) formulation – gave us access to complete pre-trained models (both encoder and decoder), suitable for task-specific fine-tuning. Finally, by scaling the model sizes from millions to billions of parameters (Brown et al., 2020; OpenAI, 2024), decoder-only LLMs have revolutionized the field by performing on par with task-specific models, substituting fine-tuning (or any other explicit training) with in-context (few-shot) learning and prompt engineering.

On the other hand, certain advances are specific to the field of Text Summarization. In contrast to other sequence-to-sequence problems, there is a major discrepancy between the length of the input and the length of the output, with the former being longer even by the order(s) of magnitude. Therefore, one of the crucial issues that is still not completely solved is the ability to encode the long input text effectively. RNNs struggled mostly with the long-term dependencies between tokens. While they are solved with each token attending to one another in the Transformer architecture (full attention), computational (memory) problems caused by the quadratic complexity emerged. Therefore, a number of solutions were proposed that compute the sparse attention only between certain tokens (see, e.g., Tay et al. (2023)).

A relevant problem concerns positional embeddings. Within the Transformer architecture, the formulation of attention requires a dedicated module that alters the token representation to consider its position (absolute or relative) in the input text. The original implementation of Transformer, i.e., Vaswani et al. (2017), explored fixed (absolute) positional embeddings that affected the performance if the length of the input was different from the average lengths encountered during training (see, e.g., Varis and Bojar (2021)). The following works proposed a modified design that considers only the relative position (distance) between tokens (see, e.g., Raffel et al. (2020); Su et al. (2024)), which can be effectively updated (see, e.g., Press et al. (2022); Chen et al. (2023)) to address the training/test input length mismatch.

Another task-specific modification to the generic architecture was proposed by See et al. (2017). The authors build upon the observation that certain entities from the input (e.g., names, dates, locations) should be preserved (copied) in the output. To allow explicit copying, the pointer-generation mechanism is implemented that computes the next-token distribution over an output vocab-

¹In this paragraph, n corresponds to the length of the input, expressed in terms of tokens.

ulary extended with a vector corresponding to input tokens. The experiments of See et al. (2017) were conducted with the LSTM variant of RNN, but in the follow-up work Enarvi et al. (2020) managed to show that the pointer-generation mechanism works also for the Transformer-based models.

A different approach was proposed by Zhang et al. (2020a), who do not modify the model but focus on the task-specific pre-training. The novel pre-training objective, Gap Sentence Generation, masks whole sentences from the input document (creating the “gap”), concatenating the gap-sentences into a pseudo-summary used as a target.

1.1.2 Extractive Summarization

In contrast to the abstractive formulation, within the extractive one, the pool of candidate summaries $\{\hat{Y}\}$ is realized directly. Namely, as summary candidates, we consider only the sub-sequences of the input document. The space of sub-sequences is limited by considering non-overlapping (ordered) input sentences, as obtained by pre-processing the input document with a sentence splitter.

By default, extractive summarization is approached as an unsupervised problem decomposed into two subtasks: sentence scoring and sentence selection. The scoring step produces sentence-level scores based on both word-level (e.g., TF-IDF importance) and sentence-level (e.g., sentence position (index) in the document, number of named entities, number of capitalized words) features. While a simple procedure that ranks the input sentences based on the scores and selects top- k scoring ones (with k depending on the desired length of the summary) as a summary is feasible, it has a number of drawbacks, e.g., similar sentences may be selected for the summary, or the distribution of scores may be biased towards certain topics (see, e.g., El-Kassas et al. (2020)). Instead, graph-based algorithms (see, e.g., Erkan and Radev (2004); Mihalcea and Tarau (2004); Barrera and Verma (2012)) are commonly applied. They transform each input sentence into a node, with edges representing the relative similarity scores. A graph-specific ranking algorithm (see, e.g., PageRank (Brin and Page, 1998)) is employed to select the final summary.

Supervised approaches to extractive summarization are limited. To obtain the reference summaries for training, the classical approach (see, e.g., Nallapati et al. (2017)) is to transform the abstractive dataset. One picks sentences from the input document that maximize a similarity metric (see Section 3.1) with the (abstractive) reference, i.e., creating the “oracle” summary. During training, the task is framed as a binary classification to decide whether a sentence should be part of the summary (sentence labeling) or not (see, e.g., Zhou et al. (2018); Al-Sabahi et al. (2018); Liu and Lapata (2019b)).

Recent approaches explore LLMs for extractive summarization by prompting the model with an explicit request to extract k input sentences as the summary (see, e.g., Zhang et al. (2023a)). However, since the final prediction is generated, we can not assure that the model will not rewrite parts of the text.

For (extractive) Text Summarization, a trivial but often strong baseline can be established by taking the first n sentences from the original document (see, e.g., Narayan et al. (2018); Lewis et al. (2020a)). This is especially true in the

news domain, where the articles are designed to catch the attention of a reader quickly and, thus, have a skewed distribution content-wise (see, e.g., [Grusky et al. \(2018\)](#)).

1.2 Vision and Language Modeling

The task of Multimodal Summarization belongs to a wider family of ViL tasks. It consists of problems that require reasoning over both textual and visual information and can be divided into a number of categories. One can consider:

- *Generative* tasks, such as:
 - Visual Question Answering ($text+image \rightarrow text$ or $text+video \rightarrow text$)
 - Image Captioning ($image \rightarrow text$)
 - Text-to-Image Generation ($text \rightarrow image$)
- *Understanding* tasks, such as:
 - Visual Entailment ($text+image \rightarrow label$)
 - Phrase Grounding ($text+image \rightarrow image\ region$)
- *Retrieval* tasks, such as:
 - Image-to-Text retrieval ($image+texts \rightarrow text$)
 - Text-to-Image retrieval ($text+images \rightarrow image$)

and many others. For a detailed taxonomy, we refer the reader to one of the recent surveys: Mogadala et al. (2021); Chen et al. (2022); Wang et al. (2022b); Gu et al. (2022); Zhou and Shimada (2023); Zhang et al. (2024).

In this section, we will focus on the technical aspect, which is common to most of those problems. Namely, we will be concerned with the problem of *multimodal encoding*, i.e., obtaining multi-modal, contextualized representation that can be passed to task-specific modules, such as the decoder module (e.g., Visual Question Answering, Multimodal Summarization) or the classification layer (e.g., Visual Entailment). We would say that the representation is *contextualized* if features from one modality had a chance to *interact* with features from another modality. The *interaction* process and how it can be modeled will be a core part of this section. The modalities on which we will focus are vision – videos (V) and images (I) – and text (T). In the following sections, we will take a closer look at the encoding process that transforms the input pair of (T, V) or (T, I) into a contextualized representation C . C will either be a single vector, i.e., $C \in \mathbb{R}^d$ or a sequence of vectors, i.e., $C \in \mathbb{R}^{P \times d}$. In Section 1.2.1, we will cover the process of embedding (feature extraction), which turns input modalities into numerical representations. In Section 1.2.2, we will cover the fusion process that contextualizes both representations into a common, multimodal representation.

1.2.1 Feature encoding

Due to the sequential nature of textual input, the size (length) of the input data can be different for every training/test sample. Additionally, due to the flexible and evolving nature of language, it is not possible to encode every word (sentence) with a fixed look-up table (see, e.g., Pennington et al. (2014); Mikolov et al. (2013)). Therefore, to encode text, we employ architectures (see Section 1.1)

capable of consuming an input of a variable length, encoding words with sequences of subword units (see, e.g., Sennrich et al. (2016); Kudo (2018); Radford et al. (2019)).

After the embedding layer (a trainable matrix, assigning a vector to every subword from the dictionary), the input text is a sequence $T = (t_1, \dots, t_K)$, with $t_i \in \mathbb{R}^{d_{text}}$. The encoding process that contextualizes the token representations (encoder) does not affect the shape of the representation. If a particular application requires a single vector to represent the whole text, two approaches are commonly used. The first one is to aggregate the vectors (along the sequence dimension) via an arithmetic sum/average/position-wise max (see, e.g., Tang et al. (2016); Cer et al. (2018)). The second one prepends (or appends) an additional [CLS] token (see Devlin et al. (2019)) to the input text and uses the contextualized representation of that token as a representation of the whole sequence (see, e.g., Radford et al. (2021)).

An image can be represented by the value of its pixels, i.e., $I \in H \times W \times N_c$, with H corresponding to the height of the image, W to its width, and the last dimension is given by the values of a color model, such as RGB ($N_c = 3$) or grayscale ($N_c = 1$). By algorithmic image scaling/resizing, one can unify a collection of images to a common, fixed input size ($H \times W$). Common sizes include, e.g., 224×224 for the ResNet (He et al., 2016) and MobileNet (Howard et al., 2017) families of models or 384×384 for the original Vision Transformer (Dosovitskiy et al., 2021). We will cover three different architectures employed as feature extractors, namely ConvNet, Vision Transformer, and Faster R-CNN (Ren et al., 2015).

ConvNet is a feedforward neural network that stacks convolutional layers with an altering kernel and stride sizes (parameters of a convolutional layer), gradually shrinking the spatial dimensions (H, W) but expanding the channel dimension (N_c) of an image, i.e., from the initial shape of, e.g., $\langle 224, 224, 3 \rangle$ to the final shape of, e.g., $\langle 7, 7, 512 \rangle$ (see, e.g., Tan and Le (2019)). Traditionally, ConvNets were trained for Image Classification (see, e.g., Deng et al. (2009)). Therefore, after the stack of convolution layers, a (global) spatial pooling operation is performed, projecting the image representation to $\langle 1, 1, d_{image} \rangle$ (the value of d_{image} is equal to the last dimension of the output, as processed with the final convolution). During training, this vector (tensor dimensions of size 1, i.e., $\langle 1, 1, \cdot \rangle$ can be squeezed) is projected to compute the value of the loss function. Otherwise, the vector gets extracted as a representation of the image for the downstream task (see, e.g., Li et al. (2020d); Fu et al. (2021); Jiang et al. (2023)). Some works (see, e.g., Zhu et al. (2018)) flatten the final spatial dimensions ($H \times W \rightarrow HW$) before the global pooling layer (e.g., $\langle 7, 7, 512 \rangle \rightarrow \langle 49, 512 \rangle$), obtaining a sequential representation of the input image. To process videos, Ji et al. (2013) proposed the 3D ConvNets that consume sequences of images/frames, performing the pooling also along the temporal dimension T , i.e., $V \in T \times H \times W \times N_c$ (see, e.g., Qiao et al. (2022)).

The architecture of Vision Transformer was proposed by Dosovitskiy et al. (2021), as inspired by the success of the textual Transformer. Vision Transformer transforms the input image $I \in H \times W \times N_c$ into a sequence p of flattened 2D patches, i.e., $p \in \mathbb{R}^{N \times (P^2 N_c)}$, where (P, P) is the resolution of each image patch,

and $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the (text-like) Transformer. By prepending a learnable embedding p_0 (similar to the [CLS] token) to the sequential representation of the image, one can either extract the encoded, sequential features $p \in \mathbb{R}^{N \times d_{image}}$ (see, e.g., Lin et al. (2023a); Zhang et al. (2022c)), or extract only the single, encoded vector $p_0 \in \mathbb{R}^{d_{image}}$ as image representation (see, e.g., Tang et al. (2024); Qiu et al. (2024)).

The Faster R-CNN (Ren et al., 2015) model is an architecture designed for object detection. The task is realized by predicting the coordinates of a rectangle enclosing the object (bounding box), along with a label corresponding to the predicted object class. On the modeling side, it first transforms the input image with a stack of convolutional layers. Then, a Region Proposal Network (RPN) component is applied to detect image regions (region proposals) likely to contain an object. Finally, after spatial pooling, the vectors corresponding to region proposals are classified into one of the object classes, with a second regression head used for bounding-box regression, i.e., smoothing the coordinates of the predicted bounding box. When used as a feature extractor, previous works identified up to k objects (hand-picked threshold), ranked according to the probability of containing an object, as computed with RPN. For each object, the corresponding feature vector (after spatial pooling, before classification/bounding-box regression) is extracted and turned into a sequence i representing an image, i.e., $i \in \mathbb{R}^{k \times d_{object}}$ (see, e.g., Li et al. (2020b)). In settings with more than one image in the input, the feature sequence is concatenated along the image dimension, i.e., $i \in \mathbb{R}^{Mk \times d_{object}}$, with M corresponding to the number of input images (see, e.g., Xiaorui (2023); Liang et al. (2023b)).

Due to the average length of a video considered in Multimodal Summarization (see Chapter 4), it is not feasible to process the whole input video with a single 3D ConvNet. Therefore, previous works have sampled either separate frames from the video (see, e.g., Li et al. (2020d); Fu et al. (2021)) or distinct subsequences of frames (see, e.g., Krubiński and Pecina (2023); Qiu et al. (2024)), that after feature extraction, were contextualized at the video-level with an RNN or Transformer-based encoder (see Section 5.1).

1.2.2 Feature fusion

In this section, we will look at the encoding process that turns the input pair of (T, V) or (T, I) into a contextualized representation $C \in \mathbb{R}^{P \times d}$. The input text T consists of token embeddings, i.e., $T \in \mathbb{R}^{K \times d_{text}}$, and the input video V is represented by a sequence of frame-level features, i.e., $V \in \mathbb{R}^{L \times d_{video}}$. While we differentiate between problems with a single and with multiple images in the input (see Chapter 2), from the perspective of feature extraction, they are similar. With multiple images in the input, we extract a vector (or a short sequence) for each input image and, thus, obtain a sequential representation. If there is only a single image in the input, it is transformed into a more expressive sequence of features, as compared to the less expressive vector² (see Section 1.2.1). Therefore, in both cases, the image input is represented by a sequence $I \in \mathbb{R}^{M \times d_{image}}$. By using a simple projection based on a dense layer (see Section 5.1), we can assume

²If it is not transformed, a single vector can be considered a sequence of length 1.

that $d_{text} = d_{video} = d_{image}$. This allows us to simplify the notation and assume that the input is a pair of textual input $T \in \mathbb{R}^{K \times d}$ and a visual input (image(s) or video) $V \in \mathbb{R}^{L \times d}$.

Modeling approaches based on RNNs contextualized the representations with the attention mechanism introduced by Bahdanau et al. (2015). The attention mechanism was first proposed for the MT task to allow the (variable length) representation in the decoder to condition on the (variable length) representation given by the encoder (see Section 1.1.1). The attention mechanism is not symmetrical, as “sequence A attending to sequence B” modifies the representation of A (not affecting the shape of A), leaving B intact. Considering our focus on the text-centric formulation of Multimodal Summarization (see Chapter 2), a majority of related works conditioned the textual representation on the visual one (see, e.g., Li et al. (2018, 2020b)), so that it could be passed to the textual decoder (see, e.g., Libovický and Helcl (2017) for a further discussion).

Within the Transformer architecture, there are two kinds of attention mechanisms: self-attention (used to contextualize, i.e., encode the input and to allow the decoder to attend to tokens generated so far) and encoder-decoder-attention (the mechanism that allows conditioning the next-token probability in the decoder based on the encoder representations). Both attention mechanisms³ are implemented with Query (W_Q), Key (W_K), and Value (W_V) matrices that project the input sequence(s). Within the self-attention formulation, all three matrices are applied to the same sequence $X \in \mathbb{R}^{N \times d}$, which gets transformed without changing its shape, i.e.,

$$Q = XW_Q, K = XW_K, V = XW_V;$$

$$X = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V \in \mathbb{R}^{N \times d}.$$

Within the encoder-decoder-attention that updates the representation in the decoder, the queries (Q) come from the previous decoder layer, i.e., $Y \in \mathbb{R}^{M \times d}$, and the keys and values from the encoder representation $X \in \mathbb{R}^{N \times d}$. This is feasible⁴, since $QK^T \cdot V = YW_Q \cdot W_K X \cdot XW_V$, i.e., $\mathbb{R}^{M \times d} \times \mathbb{R}^{d \times N} \times \mathbb{R}^{N \times d} = \mathbb{R}^{M \times N} \times \mathbb{R}^{N \times d} = \mathbb{R}^{M \times d}$.

Due to such a flexible formulation, there are a lot of ways (see, e.g., Tan and Bansal (2019); Kim et al. (2021); Xu et al. (2023b) to implement the cross-modal (text-to-vision, vision-to-text) interactions (see Figure 1.2). In principle, we differentiate between *early fusion* and *late fusion*. Early fusion (see a), b), and d) in Figure 1.2 and Section 5.2) first merges both modalities, possibly changing the sequential dimension, before computing the attention (encoding step). Late fusion first encodes each modality sequence (see c), e), and f) in Figure 1.2 and Section 5.1), before the merging step is performed. Since the encoder-decoder-attention can be stacked, it is also possible to sequentially attend to both single-modality representations and cross-modal ones, mixing in

³Please consult the original Transformer paper, i.e., Vaswani et al. (2017) for further details (e.g., Multi-Head Attention) that we simplify here for brevity.

⁴Since softmax and the normalization factor \sqrt{d} do not affect dimensions, we skip them here for simplicity.

aggregated, sentence-level or video-level features (see, e.g., Li et al. (2020c); Ging et al. (2020); Papalampidi and Lapata (2023); Xu et al. (2023a)).

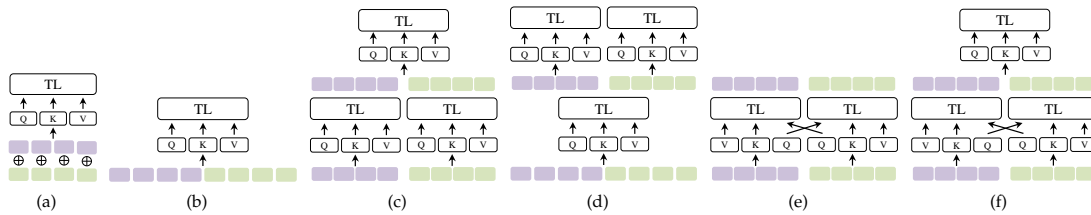


Figure 1.2: An overview of transformer-based cross-modal interactions: a) Early Summation, b) Early Concatenation, c) Hierarchical Attention (multi-stream to one-stream), d) Hierarchical Attention (one-stream to multi-stream), e) Cross-Attention, and f) Cross-Attention to Concatenation. Colors indicate features from separate modalities. Figure reprint from Xu et al. (2023a).

2. Multimodal Summarization

Parts of this chapter are based on the Ph.D. Thesis Proposal (Krubiński, 2022) submitted by the author as a part of their doctoral studies.

Following Jangra et al. (2023), we define a Multimodal Summarization task as follows:

“A summarization task that takes more than one mode of information representation (termed as modality) as input and depends on information sharing across different modalities to generate the final summary.”

Formally, let us define a *multimodal document* D_i as a tuple:

$$D_i = (M_{i1}, M_{i2}, \dots, M_{ik}) \quad (2.1)$$

where M_{ij} denotes disjoint information from a particular modality M_j , such as video (movie clip), text (textual document), or audio (voice recording) in document D_i . While using this notation, we always assume that a particular document D_i is *aligned*. By that, we mean that all modalities are coming from the same source, and the document is supposed to be presented as a whole¹ (see Figure 2.2). It might be the case that some modalities are aligned on an even finer granulation, e.g., video subtitles (text) corresponding to particular timestamps in a video clip (video). Still, we do not require it to say that the document as a whole is aligned. Therefore, the task of Multimodal Summarization can be formalized with the following formula:

$$\text{MS} : \{D_i\}_1^k \xrightarrow{\sigma} D_j \quad (2.2)$$

by which we mean the task of creating a (multimodal) summary D_j , based on a collection of input documents $\{D_i\}_1^k$ using the σ symbol to denote a summarization function. If D_j consists of a single modality (i.e., $D_j = (M_{j1})$), we talk about *Multimodal Summarization with Unimodal Output*. Otherwise, the task is called *Multimodal Summarization with Multimodal Output*.

The formula that we proposed (Eq. 2.2) is, by design, ambiguous. For example, it does not limit the *output* modalities to be a subset of *input* modalities, which is the case in the majority of applications. It also does not put any limitations on the summarization function σ – the vague definition does not enforce information sharing between the modalities. In principle, one could consider $\sigma = (\sigma_1, \dots, \sigma_k)$ such that each σ_j acts only on a single modality M_j . It would be, however, against the intended formulation, which enforces “*information sharing*”

¹A counterexample would be a multimodal document created by, e.g., combining a textual article from [Wikipedia](#) with a video obtained from [YouTube](#), and a collection of images from [Imgur](#), all retrieved with the same keyphrase. While all of the modalities will (hopefully) refer to the same event/object, they were combined artificially and were not meant to be presented as a whole.

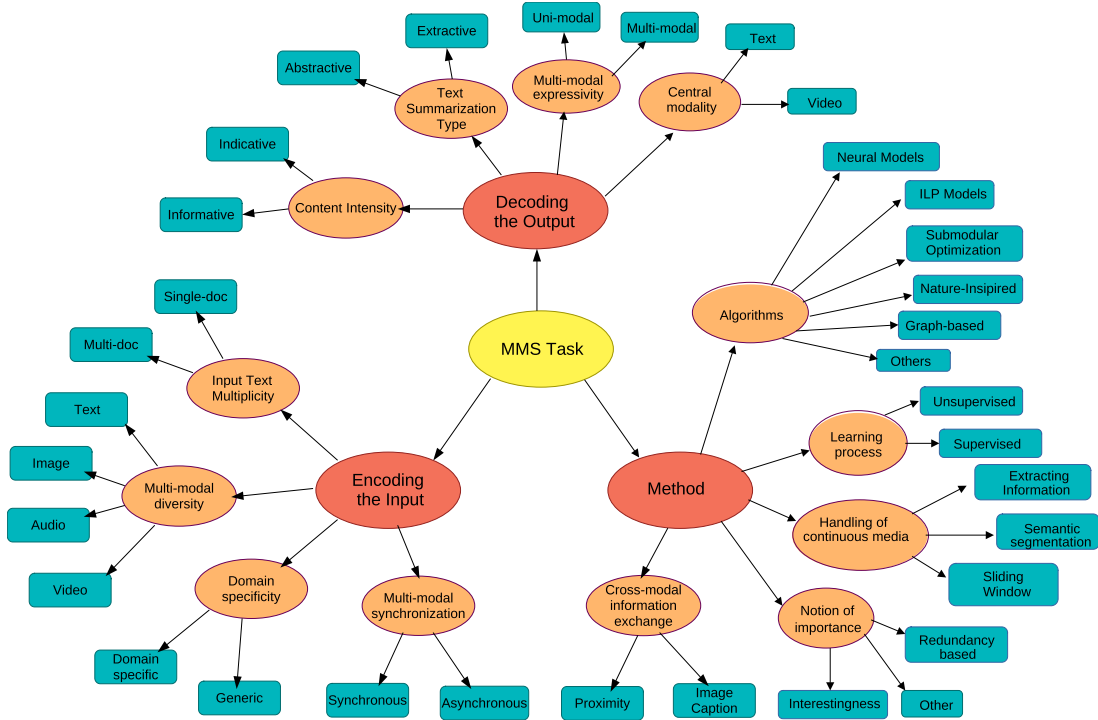


Figure 2.1: A Multimodal Summarization (MMS) taxonomy proposed by Jangra et al. (2023). The dark-orange nodes coming out of the root (in yellow) represent the segregation based on input, output, and adopted methodology. In contrast, the light-orange nodes following them represent the respective characteristics of the research work on which the works can be distinguished. The teal-colored rectangles in the leaf denote the various categories of each such characteristic. Figure reprint from Jangra et al. (2023).

across different modalities”. This ambiguity can also be noticed when comparing with the taxonomy proposed by Jangra et al. (2023) (see Figure 2.1). Therefore, we need to put in place certain limitations with regard to the Multimodal Summarization task. Unless stated otherwise, for the remainder of this thesis, we will focus on a *text-centric* Multimodal Summarization – we assume that the textual modality is always present both in the input document and in the output summary. In addition, we will be mostly concerned with the case of $k = 1$, i.e., our input will be a single multi-modal document, and our interest will mostly be targeted towards the supervised formulation. In the case of MSMO, the output modalities that we target are text plus a single image.

In the following sections, we intend to familiarize the reader with some of the particular problem variations that were approached previously. The categorization that we propose is based on the type of input/output data, as we believe it to be the most crucial one.

REVEALED: NASA's full picture set from James Webb Telescope will show detailed views of stellar nurseries with stars larger than the sun and a galaxy group 290 million light-years away


- NASA's James Webb Telescope will show new views of stellar nurseries, a galaxy group and a huge planet outside our solar system
- The space agency lists five targets for the first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT
- 'I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data,' said a longtime Webb scientist
- The release of the first images is just the beginning of Webb's scientific operations as it seeks to 'unfold the universe'

PUBLISHED: 18:24 BST, 8 July 2022 | UPDATED: 18:27 BST, 8 July 2022

NASA revealed the James Webb Telescope will target multiple spectacular cosmic objects - including far-flung stellar nurseries, a giant planet outside of our solar system and a galaxy group that's 290-million light-years away - ahead of the release of its first images.

The space agency lists five main targets for the \$10 billion telescope's first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT.

'Even after working on the program for many years, I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data from NASA's James Webb Space Telescope - an audacious endeavor in partnership with the European and Canadian space agencies,' says Eric Smith, a Webb program scientist at NASA who has been working on the telescope team since its beginnings in the mid-1990s.

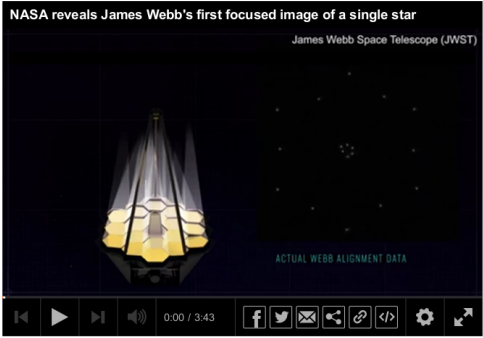


© NASA

'The James Webb Space Telescope will give us a fresh and powerful set of eyes to examine our universe,' Webb program scientist Eric Smith said. Pictured is the Carina Nebula as seen from NASA's Hubble Space Telescope

NASA reveals James Webb's first focused image of a single star

James Webb Space Telescope (JWST)



0:00 / 3:43

Figure 2.2: Example of a multimodal news article from an online publisher (dailymail.co.uk). Three modalities: text, image(s), and video are presented to a user. Each of them brings a new, unique piece of information. While particular modalities may have an inner structure – text can be split into *Title*, *Abstract*, and *Story*, in general, no specific order can be imposed on objects from different modalities.

2.1 Multimodal Summarization with Unimodal Output

In this problem formulation, the summary is unimodal and, thus, consists only of the textual modality. Therefore, one can consider purely textual baselines that generate the textual summary based only on the textual part of the input.

2.1.1 Multiple documents \rightarrow Text

Early works on Multimodal Summarization explored the usage of the secondary modalities as an auxiliary source of information to guide the refinement process of the main (textual) modality by operating on collections of *unaligned* documents. Data used in the experiments was created by manually querying a search engine for a particular phrase and collecting resources from available outputs. Tjondronegoro et al. (2011) conducted sentiment analysis of web and social media articles to annotate the key events in sports videos. Those were used as features to rank sentences from the corresponding text articles, and the top N were inputted into a pre-defined template. Li et al. (2017) collected videos and news articles² covering a hand-crafted list of recent significant world events and employed human annotators to write reference summaries.

From a modeling point of view, summaries were created in an extractive manner – non-textual features (audio features from videos transcribed to text segments, videos converted into a set of key-frames) were not used directly in the generation process but rather distilled to a set of weights. Those weights were combined with sentence-level salience scores computed with the graph-based LexRank algorithm (Erkan and Radev, 2004), and a set of sentences maximizing the objective function was considered as a summary. By ignoring sentence weights from particular modalities, the authors claimed an improvement over the text-only baseline.

2.1.2 Text + Image(s) \rightarrow Text

Li et al. (2018) introduced the multimodal sentence summarization task in the news domain, which generates a short textual summary from an <image, sentence> pair. The authors argue that the visual clues are useful for identifying the event highlights, which should help produce better sentence summaries. In their experiments, they use the <sentence, headline> tuples from the Gigaword corpus (Rush et al., 2015) and the search engine to crawl matching images. Human annotators are used to select the best-matched image for each sentence. The authors identified the need for a filtering mechanism if, e.g., the image fails to represent some abstract concepts or is too vague.

Compared to the previous works (see Section 2.1.1), the input documents are still *unaligned*, but the non-textual features are directly incorporated in the representations used for decoding. A multi-modal sequence-to-sequence model is proposed, with a bidirectional GRU (Chung et al., 2014) as input encoder and a

²In the news domain, the task of summarizing a textual document to a single sentence is usually called *title* or *headline* generation, while a longer summary is often called *abstract*. The source document is usually labeled as *article*.

uni-directional GRU, attending to both text input and image features extracted with VGGnet (Simonyan and Zisserman, 2014), as a decoder.

Li et al. (2020b) further study the filtering mechanism on the same dataset by introducing a hierarchical gating mechanism. After the encoder, hidden states are updated based on associations with the whole image (global-level gate), with image patches (grid-level gate), and with the highest scoring object proposals (object-level gate), as detected by the Faster R-CNN (Ren et al., 2015) model. An additional regularization module is proposed that uses a pairwise ranking loss to encourage a similarity score between $\langle \text{image}, \text{summary} \rangle$ to be higher than $\langle \text{image}, \text{source} \rangle$ one. The same dataset was further explored by Xiao et al. (2023), who challenged the unclear contribution of the visual modality to the quality of the final summary. They introduce a hard, binary gating mechanism that masks the whole image if the similarity between text and image features is not sufficient. Besides the negative log-likelihood loss computed between the hypothesis and reference, additional losses are introduced. They are designed to model the image complementarity for the summary by steering the generation towards tokens that are more probable based on multimodal input, as compared to those that would be generated based only on textual input. Lin et al. (2023a) argue that to capture the essence of the short, single-sentence input, it is crucial to identify *critical* tokens by exploring the visual clues from the input image.

More recent works focus on what we call *aligned* settings by collecting the data from multi-modal news websites, where the news articles are accompanied by image sequences/galleries *natively*.

Xiaorui (2023) extends the multi-lingual CrossSum dataset (Bhattacharjee et al., 2023) into the MM-CLS dataset by collecting, on average, 3.2 images per article from the website of a news provider. In their experiments, they focus mostly on cross-lingual summarization – allowed by the multi-alignment in CrossSum – and on knowledge distillation. Namely, they train a monolingual, multi-modal teacher model and separate student models for each language pair. Inspired by the approach of Li et al. (2020b), image features are extracted with a Faster R-CNN trained for object detection. Multiple images in the input are handled by simply concatenating the image vectors corresponding to the detected objects, and the architecture is based on the mT5 model (Xue et al., 2021) extended to handle visual features.

Liang et al. (2023a) build upon the MM-CLS dataset to create the M³Sum dataset by providing a further alignment that allows them to format a *cluster* of languages L_1, L_2, \dots, L_k so that the cross-lingual alignment is provided for every direction $L_i \rightarrow L_j$. This introduces a many-to-many problem, where the visual features can act as a clue for low-resource pairs. Such formulation allows a decent performance in a few-shot settings. In a similar fashion, Liang et al. (2023b) extends the XL-Sum dataset (Hasan et al., 2021) by collecting the input images from news articles. In their experiments, they provide results in high/mid/low and even zero resource settings by training on data covering almost 50 languages. The cross-lingual formulation is not considered. Similarly to Xiao et al. (2023), Faster R-CNN is explored as a feature extractor, and multiple images are handled by concatenation. Yet another contribution, the M3LS dataset, was curated by Verma et al. (2023). The M3LS dataset was derived from news articles pub-

lished by the British Broadcasting Corporation (BBC) and spans 20 languages, including a cross-lingual subset for two languages. In their experiments, Verma et al. (2023) focus on the multi-lingual aspects by exploring the mT5-based model and approach the multi-modal settings only in English, relying on the models/code by Zhu et al. (2018). The visual modality is present only on the input side, with, on average, between 1 and 5 images per article, depending on the language. The limitation with regard to the lack of a target image (multi-modal output) was observed by the authors themselves. In Section 4.3, we describe the process of collecting the image target for the subset of M3LS in English that we explored in our work on unified, multi-task, multi-modal summarization (Krubiński and Pecina, 2024).



Figure 2.3: A multimodal product summarization task proposed by Li et al. (2020a). Figure reprint from Li et al. (2020a).

Besides the news domain, a similar problem formulation was explored in the e-commerce domain, see, e.g., Chen et al. (2019); Li et al. (2020a); Rong et al. (2024). Li et al. (2020a) generate a product summary (see Figure 2.3) based on its title, description, and image, as supplied by the manufacturer. The goal of the summary is to draw the attention of a potential customer. Reference summaries were created by professionals with the goal of convincing the customer to buy the product. The proposed dataset was created based on an undisclosed e-commerce platform, and the modeling approaches are based on a sequence-to-sequence architecture with LSTM cells and cross-modal attention. Im et al. (2021) approached a similar problem, opinion summarization, in a self-supervised manner. Each instance in their dataset consists of a collection of reviews (R) describing a particular product, user-supplied product images, and additional tabulated metadata – using the data-to-text encoder proposed by Puduppully et al. (2019). A Transformer-based model is trained to generate a textual summary, using one of the reviews r_j as a target and the remaining ones R_{-j} as input. The proposed architecture is based upon the encoder-decoder text-only BART (Lewis et al., 2020a), which is firstly pre-trained for opinion summarization on textual

data. Image features are extracted with the ResNet101 (He et al., 2016) model. In the decoder, the attentions to each input modality (text, image, metadata) are combined via a weighted, position-wise addition. This setup allows the authors to analyze the influence of each modality and shows that the multi-modal clues are not very effective (e.g., ROUGE-L³ 19.84 \rightarrow 19.54 without the image modality).

Recent works keep expanding the task to new domains. Overbay et al. (2023) proposed the mRedditSum dataset (3,030 instances) based on discussion threads from Reddit. By manually filtering subreddits with discussions surrounding visual content, authors were able to create a dataset in which each instance consists of a single image, with the textual modality provided by the content of the post from the original uploader and comments from other users. The summaries were written by workers from a crowdsourcing platform.

2.1.3 Text + Video \rightarrow Text

Besides the image modality, a number of works explored the summarization problem that generated the textual summary based on a textual document and a (relatively) short video.

The first large-scale resource that facilitated such research – the How2 dataset – was introduced by Sanabria et al. (2018). The authors collected almost 80,000 instructional videos from the YouTube platform, totaling roughly 2,000 hours. For each video, corresponding English subtitles and video descriptions were collected. While the video descriptions were manually written by the video creators, there is no guarantee that automatic ASR systems were not involved in creating subtitles to some degree. In their experiments, the authors propose a summarization task that, based on subtitles and video frames, generates the description (summary). A 3D ConvNet model, namely the ResNeXt-101 3D (Hara et al., 2018) trained for action recognition on the Kinetics dataset (Kay et al., 2017), is explored as a feature extractor, generating a single vector for every 16 frames. Text – both input and output – is lowercased and tokenized. Audio features (43-dimensional vector for every window of 25 milliseconds) are extracted with Kaldi (Povey et al., 2011). Due to the copyright laws (see Chapter 4), the dataset is shared only as those pre-computed features, forcing the same feature encoding/processing from all follow-up works. Palaskar et al. (2019) proposed a RNN-based modeling approach that builds upon the findings on combining attention in multi-encoder setup by Libovický and Helcl (2017). The best multi-modal model they train is only marginally better than a SOTA text-only model (ROUGE-L of 54.9 vs 53.9). Khullar and Arora (2020) explored the addition of audio modality via a trimodal, hierarchical attention mechanism, i.e., text features attend independently to audio and video features, in the next step, combining those into a unified representation. While they report an improvement over the text-only baseline (ROUGE-L of 42.23 vs 39.98), the number of trainable parameters greatly differs, raising questions about the validity of the results. The audio-video-text model has 32.08M parameters vs 16.95M of the video-text model – the number for the text-only variant is not reported.

More recent works switched to the Transformer architecture. Liu et al. (2020)

³Please consult Section 3.1 for a thorough discussion on evaluating the quality of textual output. Unless stated otherwise, the F1 variant of ROUGE is reported.

introduced the Forget Gate mechanism (see Section 5.1), which allows the model to dynamically control the information from auxiliary modalities by scaling them down. The authors also explore an off-the-shelf ASR system to generate input text (subtitles) instead of relying on those provided by Sanabria et al. (2018). When substituting the original subtitles with the ones generated by the ASR system, a much smaller drop in performance of the multi-modal model (ROUGE-L 58.2→56.1) compared to the text-only one (ROUGE-L 53.8→43.3) suggests that the usefulness of visual features is much higher if the core, textual content is of lower quality. This is in line with our findings related to the pre-training, as presented in Section 5.1. Yu et al. (2021) explored the effective ways of fusing visual features with pre-trained, text-only models. Namely, they explore two cross-attention formulations (a simple one based on video-to-text dot-product attention and one based on multi-head formulation with the cross-attention formulation from classic decoder) and try to experimentally determine which layer and which model component (encoder vs decoder) are best suited to perform the cross-modal attention. The findings suggest that cross-modal attention should be performed in the encoder and at higher layers. The numerical improvements attributed to the visual modality are marginal (ROUGE-L 57.5→58.0 for a variant based on T5) but improve once the Forget Gate and additional encoder that contextualizes visual representations are employed (ROUGE-L 61.4→64.4 for a variant based on BART). By substituting video features with random noise during the inference, the authors claim the robustness of the text-only component. However, this also proves that the visual features are not effectively consumed. The independent experiments by Xu et al. (2023d) came to the same conclusion that the higher encoder layers are the optimal place to implement cross-modal attention. Thanks to the work of Sanabria et al. (2018), who translated the source text (subtitles) from the test-set into Portuguese, some works, e.g., Liu et al. (2022a), explored the cross-lingual settings, but this research direction did not catch a lot of attention.

Besides the How2 dataset, a number of other resources were introduced. However, since they were not publicly shared, most of them were explored only in a single work. Qiao et al. (2022) propose a WB-News dataset in Chinese based on the Weibo social media platform. In their experiments, they follow Yu et al. (2021) and focus on effective ways of fusing visual features with pre-trained text-only models. They also highlight the importance of task-specific pre-training that aligns the visual and textual features. Still, the benefits of including the visual information in the input are not transparent – even during human evaluation, the multi-modal variant is only marginally better. Using a Likert scale (of 1-5), the best multi-modal variant scores on average 3.71, while the best text-only model achieves an average score of 3.65. Faheem et al. (2024) introduced the first multi-modal dataset in Urdu, targeting Urdu news channels on YouTube. In their experiments, they follow the MMS architecture (without the image decoder) proposed by Krubiński and Pecina (2023). The video-based formulation was also recently approached by Tiwari et al. (2024) in the medical domain, with the creation of the MM-MediConSummation dataset. The MM-MediConSummation dataset consists of 467 audio/video recordings of doctor-patient counseling sessions that were manually annotated by medical graduate students. Transcripts and textual summaries were collected, with the guideline to focus on specific

aspects, such as the gender of the patient, their age, or the primary intent for consultation. In their experiments, the authors focus on the benefits of including non-textual modalities in the input, reaching a conclusion that they are beneficial to the final quality. A crucial aspect related to the anonymity of patients is not discussed, with the authors exploring even patient-specific features such as age or gender.

2.1.4 Other formulations

Besides the textual and visual (images, videos) modalities, the audio modality was also explored (see, e.g., [Sharma et al. \(2022\)](#); [Jung et al. \(2024\)](#)). However, we are not aware of any work that would use the audio modality to extract *independent* information. If the audio modality consists only of speech, we believe that the same content can be expressed as text by applying an [ASR](#) system. A certain amount of information that could be beneficial for summarization may be expressed only via the audio modality, e.g., urban noises suggesting that the audio was recorded in a city or animal noises suggesting that the recorded event/conversation took place at a farm, but this research direction has not yet been explored.

A number of other modalities/formulations were also explored to generate textual summaries. [Chen et al. \(2019\)](#) adds user-specific information (gender, age, etc.) and additional clues extracted from a knowledge graph to generate product descriptions in the e-commerce domain. [Trieu et al. \(2020\)](#) generates textual descriptions based on a coherent set of images. This task is similar to image captioning, but by aggregating different photos/instances of the same object, it converges toward summarization. [Himakunthala et al. \(2023\)](#) generate video descriptions in a step-by-step manner. The authors start by identifying keyframes, which are turned into a sequence of sentences (by image captioning) and image features (by an object detection model). In the next steps, both of those sequences are consumed to generate an unstructured, dense description. Finally, a [LLM](#) is applied to turn those into a structured, template-like summary with desired properties.

2.2 Multimodal Summarization with Multimodal Output

In this problem formulation, the summary is multimodal and consists of a text accompanied by a different modality. As announced in Chapter 2, our core interest lies in formulations in which the output (summary) consists of text and a single image.

2.2.1 Text + Images \rightarrow Text + Image

The image-based formulation of MSMO was introduced by Zhu et al. (2018). In their foundational work, the authors curate a novel dataset based on the Daily Mail portal, collecting the textual articles (source documents and target summaries) together with the input images (on average, 6.6 images per article), which are presented to the readers to enrich the textual information. The pictorial reference is not a native part of the data – the authors employ graduate students to pick up to three relevant images per article, annotating only the testing part of the data. Thus, during training, there is no direct supervision from the visual modality. Instead, in their experiments (based on a sequence-to-sequence architecture with LSTM cells), the authors include a coverage loss that encourages higher values of attention in the cross-modal attention block. During decoding (inference), the coverage vector (text-to-image attention) is used to choose the most relevant vector. This builds upon the intuition that the most relevant image should be the one with the highest values of attention, i.e., the one deemed the most useful by the textual decoder. Human evaluation is performed to judge the quality and usefulness of pictorial summaries. The authors measure the “satisfaction of informativeness”, i.e., whether the annotators consider that the image contributed positively to their understanding of the summary. The results suggest that people might prefer multimodal summaries over text-only ones. When evaluating the quality of textual summaries with automatic metrics, the best multi-modal model is only marginally better than the best text-only one (ROUGE-L of 37.74 vs 37.75).

In the follow-up work (Zhu et al., 2020), the authors propose a method to incorporate direct guidance of visual modality during training. Namely, they impose an absolute ordering of input images – either based on their position in original news articles (OR) or based on the semantic similarity between the reference summary and the caption collected for each image (RR). Next, they arbitrarily choose a value k and treat the top- k images as the target and the remaining ones as negatives. This information gets incorporated into the training as a classification task that, during inference, computes the similarity between the representation of a particular image and either the encoded source (ENC) or the summary generated by the decoder (DEC). The authors conclude that the RR/DEC variant performs the best, although the findings differ based on which automatic metric is used.

Zhang et al. (2022c) approach the multi-task formulation by generating both the extractive and abstraction summary. The gold-standard extractive summaries required for training are obtained with an oracle approach that maximizes the ROUGE score with the original reference. Jiang et al. (2023) use only the textual

modality (summaries) as a training signal, exploring pseudo captions to choose the most relevant image. Namely, for each input image, they explore a text-image alignment and employ a retrieval model to pick a sub-sequence of tokens from the gold-standard summary that acts as a pseudo caption for the image. Those captions are used to train a task-specific image captioning model required for inference. Finally, during inference, once the textual summary is generated, an image with the highest lexical similarity between the generated summary and its pseudo caption is chosen as the pictorial summary.

More recent works approach the problem in a fully supervised manner. [Zhang et al. \(2022a\)](#) propose a Chinese dataset based on the text-only TTNews ([Hua et al., 2018](#)) and THUCNews ([Sun et al., 2016](#)) datasets. They use a search engine to collect up to 10 relevant images (with corresponding captions) for each article. In the next step, a semantic matching model is used to filter up to three most relevant images per article. Finally, human annotators are employed to pick the most relevant image out of those three. This setup allows training with the image supervision directly – given a sequence of images I , one of them has a positive “relevance” label and all the other negative one. During training, image features are projected and transformed to a numerical score $s_{i \in I} \in \mathbb{R}$, which, combined with negative log-softmax, allows computing the log-likelihood loss with respect to “relevance” annotations. [Zhang et al. \(2023b\)](#) explore the same dataset and introduce auxiliary losses with the goal of enhancing multimodal semantic coverage. Specifically, they employ an additional textual decoder to generate a visual description for each input image. Compared to simple image captioning, those descriptions have access to the background information provided by the input article. The description of the most relevant image, as deemed by the model, is concatenated with the textual summary and, together with the image itself, forms a multimodal output.

Similarly to the formulation with text-only output (see Section 2.1), the majority of the works consider the multimodal summarization problem in the news domain. Works in the other domains exist but are much more scarce. [Rong et al. \(2024\)](#) use the dataset of multimodal product descriptions (textual descriptions with multiple images) from a Chinese e-commerce platform introduced by [Li et al. \(2020a\)](#). The authors claim to have access to a version of the dataset with annotated target images, which were not mentioned in the original paper. Since neither the dataset nor the code is publicly available, we are unable to confirm the validity of the setup. On the modeling side, the authors propose a novel approach that employs visual information to explicitly modify the distribution of possible words. Instead of extending the target vocabulary with source tokens, similarly to the Pointer-Generator network (see Section 1.1.1), the authors instead try to limit the vocabulary by disregarding words that the model deems irrelevant, given the multi-modal context. Cross-Entropy loss is computed based on the projected, single-dimensional, and contextualized image representations and the gold-standard annotations, directly incorporating the visual signal during training.

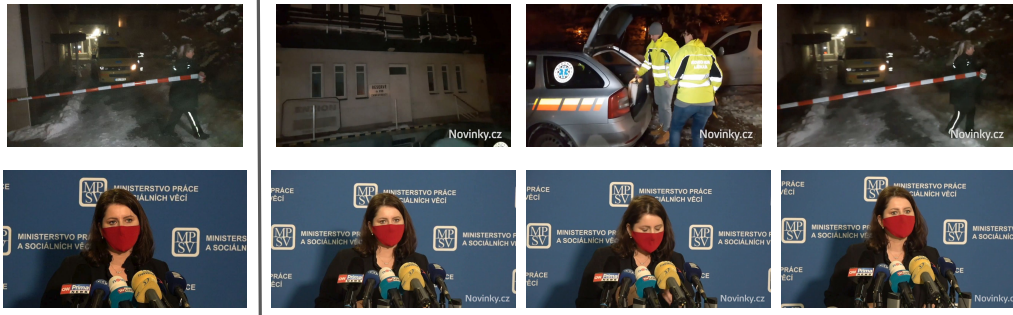


Figure 2.4: Examples of visual modality from the MLASK (Krubíňski and Pecina, 2023) dataset. Left – the target image. Right – a subset of input video frames, as seen by the model. The target image was modified by removing the watermark in the bottom-right corner.

2.2.2 Text + Video \rightarrow Text + Image

The video-based MSMO formulation (VMSMO) was first proposed by Li et al. (2020d). Despite the fact that, for technical reasons, the video is typically converted to a sequence of (down-sampled) frames and thus can be treated as a collection of images, there are some crucial differences compared with the image-based formulation. Firstly, the temporal dependency between the frames is clear and well-defined. This is not the case for inputs consisting of multiple images, which are commonly encoded independently of one another. Secondly, within the image-based formulation, it is assumed that the target image (pictorial summary) is one of the input images. In the video-based formulation, one assumes that the image target is *very similar* to one of the frames. The image target may be created by applying minimal edits, such as cropping or watermark removal. It may also happen that due to the frame down-sampling, the *exact* frame, which resembles the pictorial summary (i.e., the case when the image target is indeed a frame from the video), gets dropped (see Figure 2.4 for an example from the MLASK (Krubíňski and Pecina, 2023) dataset).

In their field-defining work, Li et al. (2020d) argue that in real-world applications, a text article is usually accompanied by a video consisting of hundreds of frames rather than a few images. Therefore, they propose to choose a single frame to act as a pictorial summary that should represent the salient point of the whole video. To facilitate their research, they collect a dataset from the largest social network website in China. Besides individuals, China’s mainstream media also have accounts on that platform, which they use to post short, lively videos and articles. Each instance in the curated dataset contains a textual article, a textual summary, and a video with a reference cover picture. In their experiments, the cover picture is not used directly. Instead, they regard the frame that has the maximum cosine similarity with the reference cover picture as the positive sample and all the others as negative samples. The authors report that the average cosine similarity of the positive sample is 0.90, proving the validity of the setup. In their experiments, they employ a dual encoder setup with separate encoding blocks for videos and text. Video features are extracted only at the frame (image) level, with a RNN network employed to contextualize the representations. The supervision from the visual modality is provided by a pairwise hinge loss that

awards assigning a higher matching score (computed based on the contextualized frame representations) to the positive sample (one with the highest similarity to the target image), as opposed to the remaining ones from the input video. Video pre-processing is done by extracting one of every 120 frames, aiming to obtain roughly 10 candidate frames. Evaluation with automatic metrics suggests that the multimodal variant generates the highest quality of textual summaries, with similar findings based on human evaluation. Yet, the authors train their models from scratch without exploring even a pre-trained text encoder and compare only to self-trained models without including the results on any well-established textual benchmark, raising questions about the validity of the setup.

In a follow-up work, [Fu et al. \(2021\)](#) present a full-scale multimodal dataset comprehensively gathering documents, summaries, images, captions, videos, audios, transcripts, and titles. The dataset was collected from well-known English news websites, namely [CNN](#) and [Daily Mail](#). Compared to [Li et al. \(2020d\)](#), the proposed dataset does not include a single reference picture (images are part of the input document) and thus utilizes unsupervised methods during training⁴. The authors still report the aggregated, average cosine similarity between the related images and the frame selected by the model. It should be highlighted that the best-performing variant does not achieve significantly higher average cosine similarity than a trivial baseline that picks a random frame from the video – 69.22 vs 67.69. On the modeling side, the architecture utilizes [RNN](#) encoder-decoder setup with bi-directional LSTM cells, and the video encoding is done at the frame level, with an encoder providing the temporal dependencies. Our work (see Section 5.1) based on the MLASK dataset ([Krubiński and Pecina, 2023](#)) was one of the first formulations of [VMSMO](#) that incorporated the frame/image similarity directly into the training and employed the 3D [ConvNet](#) to model the temporal dependency between consecutive frames.

[Tang et al. \(2024\)](#) proposed the extreme TL;DW (“Too long; didn’t watch”) formulation of [VMSMO](#) that summarizes a pair of text and video to a single frame and a single sentence. The authors curate a novel resource based on [YouTube](#) and devise a new unsupervised training strategy based on the optimal transport theory. One of the losses incorporated during training assures the cross-modal similarity of the output by maximizing the text-image similarity measured with CLIP ([Radford et al., 2021](#)). The video thumbnails act as cover pictures, and the pixel-level Euclidean distance is computed between the reference and the frame picked by the model during evaluation. Comparatively low values of automatic metrics – ROUGE-L of 4.33 achieved by the best model – make it difficult to analyze the final results.

A recent work of [Qiu et al. \(2024\)](#) builds upon the architecture and formulation proposed in the MLASK ([Krubiński and Pecina, 2023](#)) paper by incorporating separate frame- and video-level encoders. In their unique formulation, the textual summary is inpainted onto the cover frame to simulate the video thumbnails from [YouTube](#). In their experiments, the authors approach the challenging problem of

⁴The version of the work accepted to the ACL Anthology – [Fu et al. \(2021\)](#) – provided the results only for the textual output. However, the original, extended version of the work published on arXiv – [Fu et al. \(2020\)](#) – also provided the results for visual output. For consistency, we refer to the version that got accepted to a venue and officially published, but also refer to the results from the preprint.

temporal segmentation. A vast majority of previous works sampled the frames uniformly by skipping all but every n -th frame. In contrast, Qiu et al. (2024) argue that an initial step of segmenting the video into a sequence of scenes, followed by scene-level frame sampling, is crucial. We approach this strategy in Section 5.2.4, with the results of our experiments suggesting that the frame-sampling step may indeed be of significant importance to the overall quality of the pictorial summary.

2.2.3 Other formulations

Besides the MSMO formulations covered in previous sections, other variants that output the $\langle \text{text}, \text{image} \rangle$ pair exist but are by far less prevalent.

The first one we would like to cover is the “Video \rightarrow Text + Image” formulation, a subtask of the more generic cross-modal video summarization problem. The idea here is to extend a video summarization model with a text decoder that generates a short summary based on the frame- or video-level representations, similarly to the Video Captioning problem. The summary can be generated either based on the whole video (see, e.g., Lin et al. (2023b)) or in a hierarchical manner that aggregates the partial clues from the frame- or segment-level representations (see, e.g., Papalampidi and Lapata (2023)).

The second one is the “Text \rightarrow Text + Image” formulation⁵, enabled by the progress in generative AI and fueled mostly by the Generative Adversarial Nets (Goodfellow et al., 2014) and Stable Diffusion (Rombach et al., 2022) approaches. We are not aware of any work that would train a single (summarization) model to generate both the text and the image – previous approaches explored two separate text-to-text and text-to-image models, see, e.g., El et al. (2019) or our work (Krubiński and Pecina, 2024) on unified Multimodal Summarization in Section 5.2.

⁵With the recent progress in video generation, see, e.g., Sun et al. (2024), a MSMO formulation with $\langle \text{text}, \text{video} \rangle$ output seems to be a natural next step.

3. Quality Evaluation

In Chapter 1, we discussed the difference between *extractive* and *abstractive* summarization. If we limit ourselves to the supervised settings, then the extractive summarization problem can¹ be treated as a task of binary classification, thus allowing us to work with the classical Accuracy/Precision/Recall/F-measure metrics. If we frame the problem differently, and instead of predicting whether a sentence/frame is relevant (binary label), we predict *how* relevant it is, the problem can be treated as either a regression or retrieval task. It should be noted that a regression problem can be converted to a (binary) classification problem by using thresholds to group scores into classes. Similarly, a number of Machine Learning algorithms (including neural networks) do not predict a class directly but rather predict a distribution over classes, i.e., a probability of belonging to a class. If we treat the problem as a regression, one can compute the prediction error directly via, e.g., Mean Absolute Error (MAE), Mean Squared Error, or Root Mean Squared Error. If we approach it as a retrieval task (How many of the sentences/frames, in summary, are relevant? What proportion of relevant sentences/frames did we choose? Were the sentences/frames chosen for the summary *the most* relevant?), then one should report a subset of retrieval metrics best suited for the particular application (see Section 3.2).

The situation becomes much more challenging if we frame the problem in an *abstractive* manner, generating the summary from scratch – based on the input representation, but without directly copying parts of the input. In that case, we do not have a fundamental measure telling us how similar two sentences (or two images²) are. Rather, we fall back on using either human annotators (see Section 3.1.1 and Section 3.3) or using an automatic heuristic.

In the following sections, we intend to provide an overview of metrics and evaluation protocols used to evaluate the quality of textual output (see Section 3.1), visual output (see Section 3.2) and, finally, multimodal output (see Section 3.3).

¹Assuming that we make some additional assumptions, such as taking the sentence splitting/frame sampling for granted, or fixing sentence/frame ordering.

²One could argue that if two images have the same size, then comparing the values of RGB pixels could work. It was shown, however, to be ineffective. The value of a difference would be large for small deformations (e.g., the distance between an image and the same image shifted by one pixel), and, overall, is sensitive to noise and changes in brightness, see, e.g., [Nakhmani and Tannenbaum \(2013\)](#).

3.1 Textual Output

The task we are concerned with is as follows: given a textual document D , a reference summary S , and a summary Y (of D) generated by an automatic system Sys , automatically assign a numerical score $m = M(D, S, Y)$ that measures the *quality* of Y . One differentiates between *reference-based* and *reference-free* evaluation protocols. If one has access to the reference summary S , i.e., $m = M(D, S, Y)$ or $m = M(S, Y)$, the task is called *quality evaluation*. If the reference summary S is not available (unsupervised settings), the task is called *quality estimation*, i.e., $m = M(D, Y)$.

In practical applications, we are mostly concerned with the performance not on a single document D_i , but rather on a whole test-set: $\{D_i\}_1^K$. The score m assigned to the whole test-set is typically based on aggregated document-level scores m_i , usually by an arithmetical average, i.e., $m = \frac{1}{K} \sum_1^K m_i(D_i, S_i, Y_i)$, but can also incorporate some corpus-level statistics, e.g., a frequency of words/terms. For the remainder of this chapter, unless stated otherwise, the default aggregation method will be the average. Thus, we will be discussing metrics at the document-level, as the extension to the corpus-level is trivial. Since the numerical scores m that we commonly operate with are corpus-level based, it is considered a bad practice to compare scores computed based on two (or more) distinct test-sets. Similarly, given two metrics M_1 and M_2 , we can not make any assumptions regarding the distribution of scores that they generate (see, e.g., Kocmi et al. (2024)). Therefore, it should be noted that the intended usage of automatic metrics is to answer the following question: “Given test-set $\{D_i\}_1^K$, and two automatic systems Sys_1 and Sys_2 , decide which performs better, according to automatic metric M ”, which often gets extended³ to: “Given test-set $\{D_i\}_1^K$, and a number of automatic systems $\{Sys_j\}_1^L$, find a system Sys_j that performs best, according to automatic metric M ”.

The remaining question concerns the notion of “*quality*”. As discussed, automatic metrics are heuristics that one explores when human annotation, which should be considered the gold standard, is not feasible. In Section 3.1.1, we will outline several methodologies for the annotation process. For now, let us assume that for a given test-set $\{D_i\}_1^K$ a human annotation was conducted and that each summary Y_{ij} (Y_{ij} stands for a summary of document D_i , generated by system Sys_j) had been assigned a score $h_{ij} \in \mathbb{R}$. Given such scores h_{ij} , how can we decide which of the metrics M_1, M_2, \dots, M_q should be picked in order to decide which system Sys_j performs best?⁴ The commonly accepted answer is to pick the metric M_i that achieves the highest correlation coefficient $Corr(\cdot)$ with human scores. Commonly explored coefficients include Pearson’s r , Spearman’s ρ , and Kendall’s τ (see, e.g., Deutsch et al. (2022)). Most of the works report either *segment-level*⁵ correlation, i.e., $Corr(\{h_{ij}, m_{ij}\}_{i=1, j=1}^{K,L})$, or *system-level* correlation, i.e., $Corr(\{\frac{1}{K} \sum_{i=1}^K h_{ij}, \frac{1}{K} \sum_{i=1}^K m_{ij}\}_{j=1}^L)$. An alternative approach that computes pairwise accuracy was recently proposed by Kocmi et al. (2021), which, instead of relying on correlation, explicitly counts the number of pairwise rank

³Some works argue that this logic may be flawed, see, e.g., Kocmi et al. (2021)

⁴It may happen, that according to metric M_1 system Sys_3 performs best, but according to metric M_2 system Sys_5 is better.

⁵Following the no-grouping formulation, see Deutsch et al. (2023) for an extensive discussion.

agreements between metric and human scores. Given (average) human scores h_i and (average) metric scores m_i for system Sys_i , we count the number of system pairs $\{Sys_i, Sys_j\}$ for which both human annotators and the automatic metric agree on the relative performance, e.g., both concluding that Sys_i is better than Sys_j , i.e., $m(Sys_i) = m_i > m_j = m(Sys_j)$ and $h_i > h_j$.

3.1.1 Human Evaluation

In the related field of Machine Translation, thanks to the Metrics Shared Task (Freitag et al., 2023, 2022, 2021; Mathur et al., 2020) collocated with the WMT (Conference on Machine Translation, historically Workshop on Statistical Machine Translation) conference since 2008 (Callison-Burch et al., 2008), advances in the MT models performance are accompanied by continuous development of new automatic metrics that improve correlation with human judgment and are robust to both domain shifts and changes in annotation style. Thanks to such a long-lasting⁶ and centralized initiative, the annotation methodologies for MT are standardized and generally produce a single numerical score that measures the overall quality of translation.

However, this is not the case for Text Summarization. A number of independent works analyzed the automatic metrics by comparing them with human judgments (see, e.g., Bhandari et al. (2020); Fabbri et al. (2021); Liu et al. (2023a); Deutsch et al. (2022, 2023)). Since there is only a partial overlap in terms of metrics, datasets, and annotation methods that were examined, a fair meta-evaluation is not possible. A majority of recent works that conduct manual evaluation label the summaries based on a number of (mostly independent) *dimensions* (see Figure 3.1). Koto et al. (2022) identified four key dimensions across which to evaluate summaries, namely: faithfulness, focus, coverage (see Figure 3.2) and inter-sentential coherence (also named fluency). They define them as follows:

- **Faithfulness** – the degree of factual consistency (and lack of hallucination) with respect to the source article;
- **Focus** – assesses semantic equivalence by evaluating the proportion of important information in the generated summary (precision);
- **Cover** – assesses semantic equivalence by evaluating the degree of salient information in the reference summary that the generated summary contains (recall);
- **Fluency** – the degree to which the summary sounds natural and has no grammatical problems.

A different set of dimensions were proposed by Fabbri et al. (2021):

- **Coherence** – the collective quality of all sentences, as in “The summary should be well-structured and well-organized. The summary should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.”

⁶In 2001, a first edition of Document Understanding Conference (initially Workshop on Text Summarization) took place. Similarly to WMT, it was centered around constrained settings explored via shared tasks. However, the event was discontinued in 2007 (<https://duc.nist.gov/pubs.html>). Thus, it did not have a chance to establish itself and propagate standardized protocols properly.

Paper	Automatic					No manual eval	Manual												
	ROUGE	METEOR	BLEU	BERTScore	MoverScore		Faithfulness	Recall	Precision	Relevance	Coherence	Fluency	Relative	Absolute	SCU	reference	article	ref+article	Quality control
See et al. (2017)	✓	✓				✓													
Yang et al. (2017)	✓						✓					✓			✓				
Lin et al. (2018)	✓					✓													
Cohan et al. (2018)	✓					✓													
Liao et al. (2018)	✓					✓													
Kedzie et al. (2018)	✓					✓													
Amplayo et al. (2018)	✓					✓			✓		✓	✓				✓			
Jadhav and Rajan (2018)	✓					✓													
Li et al. (2018a)	✓					✓													
Pasunuru and Bansal (2018)	✓	✓				✓													
Cao et al. (2018)	✓					✓													
Sakaue et al. (2018)	✓					✓													
Celikyilmaz et al. (2018)	✓					✓	✓	✓		✓	✓	✓	✓			✓			
Chen and Bansal (2018)	✓	✓				✓			✓	✓	✓	✓	✓				✓		
Guo et al. (2018)	✓	✓				✓			✓	✓	✓	✓	✓				✓		
Hardy and Vlachos (2018)	✓		✓			✓				✓		✓	✓						
Hsu et al. (2018)	✓					✓	✓	✓		✓		✓	✓			✓			✓
Krishna and Srinivasan (2018)	✓					✓	✓			✓		✓	✓			✓			✓
Kryściński et al. (2018)	✓					✓				✓		✓	✓			✓			✓
Li et al. (2018b)	✓					✓	✓					✓	✓			✓			✓
Narayan et al. (2018a)	✓					✓	✓				✓	✓	✓			✓			✓
Narayan et al. (2018b)	✓					✓	✓				✓	✓	✓			✓			✓
Narayan et al. (2018c)	✓					✓	✓				✓	✓	✓			✓			✓
Peyrard and Gurevych (2018)	✓					✓	✓	✓			✓	✓	✓		✓				✓
ShafeiBavani et al. (2018)	✓					✓	✓		✓	✓	✓	✓	✓	✓					✓
Song et al. (2018)	✓					✓	✓				✓	✓	✓			✓			✓
Hardy et al. (2019)	✓					✓	✓	✓			✓	✓	✓			✓			✓
Makino et al. (2019)	✓					✓													

Figure 3.1: An overview of evaluation methods by Koto et al. (2022). In the “Manual” column, we see the annotations related to the dimension of human evaluation (Faithfulness, Recall, etc.), a label telling us whether the evaluation was conducted on individual documents (Absolute) or in the context of other texts (Relative), and a taxonomy of methodologies for human evaluation. Figure reprint from Koto et al. (2022)

- **Consistency** – the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document.
- **Fluency** – the quality of individual sentences, as in “The summary should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.”
- **Relevance** – to measure the selection of important content from the source. The summary should include only important information from the source document.

Those categories differ from the ones proposed by Harman and Over (2002) in the pioneer evaluation campaign conducted as a part of the first Document Understanding Conference (DUC) in 2001:

- **Grammaticality** – [All, most, some, hardly any, or none] of the syntactic units (e.g., sentences, clauses, phrases, etc.) follow the rules of English

grammatical form (independent of content).

- **Cohesion** – [All, most, some, hardly any, or none] of the sentences fit in as they should with the surrounding sentences.
- **Organization/coherence** – [All, most, some, hardly any, or none] of the summary is well-organized, i.e., the content is expressed and arranged in an effective way.

Gold summary : Info-A; Info-B; Info-C
System summary:
Good <i>focus</i> , and Good <i>coverage</i> : Info-A; Info-B; Info-C
Good <i>focus</i> , and Bad <i>coverage</i> : Info-A; Info-A
Bad <i>focus</i> , and Good <i>coverage</i> : Info-A; Info-B; Info-C; Info-D; Info-E
Bad <i>focus</i> , and Bad <i>coverage</i> : Info-D; Info-E; Info-F

Figure 3.2: Illustration of *focus* and *coverage* (see Section 3.1.1). Figure reprint from Koto et al. (2022)

Other works have either proposed a direct mapping between certain dimensions (e.g., *Focus* unified with *Relevance*), explored a more fine-grained taxonomy (e.g., the types of errors proposed for factual evaluation by Huang et al. (2020) are: *Addition*, *Omission*, *Inaccuracy intrinsic*, *Inaccuracy extrinsic*, *Positive-negative aspect*), or defined the evaluation dimensions themselves (e.g., *Overall Quality* dimension in the annotation campaign organized by Stiennon et al. (2020)).

Having discussed the *evaluation dimensions*, we will now focus on the *evaluation methodologies* for the annotation process.

How much information contained in the black text can also be found in the gray text?

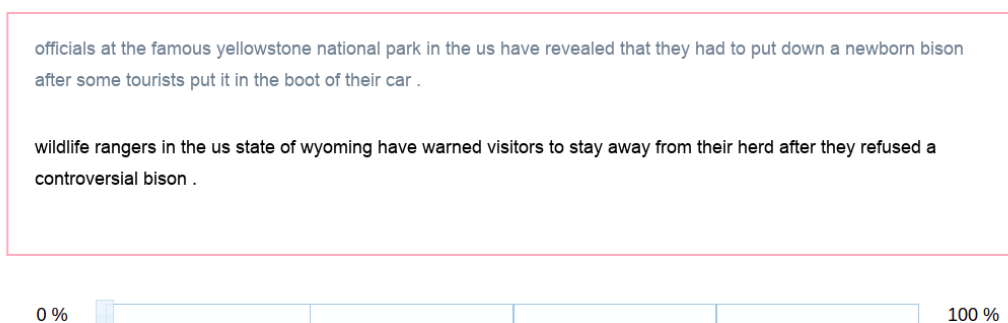


Figure 3.3: Illustration of the Direct Assessment evaluation framework. By switching the position of reference and hypothesis, the same question can be used to collect annotations for *focus* and *coverage*. Figure reprint from Koto et al. (2022)

- **Direct Assessment** – The Direct Assessment (DA) framework was first proposed by Graham et al. (2015) to evaluate quality of MT outputs. Until 2021, it was used as a standard evaluation methodology in WMT. In the context of summarization, this methodology was explored by, e.g., Koto et al. (2022) or Koto et al. (2021). During the annotation process, a human is shown a question and two pieces of text – usually the reference summary and a summary generated by an automatic system (see Figure 3.3). The

task is to choose a number between 0 and 100, which will act as the human score h . To make the process more robust, after the annotation process is completed, scores from each annotator are usually normalized (across all documents) to a z-score by subtracting the mean and dividing by standard deviation.

- **Likert Scale** – The Likert Scale (Likert, 1932) has its roots in psychological research. When responding to a Likert item, annotators specify their level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements, e.g., choosing 5 for “Strongly agree” and 1 for “Strongly disagree” (see Figure 3.4). This framework was explored by, e.g., Fabbri et al. (2021) or Stiennon et al. (2020). The advantage it has over DA is that it allows the researchers who design the annotation process to explicitly describe the expected quality of each item on the scale.
- **Content-based** – Nenkova and Passonneau (2004) designed the Pyramid protocol specifically for the purpose of summary evaluation. It was inspired by the early works probed within the Document Understanding Conference in 2003. It does not require humans to directly judge the quality of a summary on a numerical (or ordinal) scale. Instead, the process starts by identifying Summarization Content Units (SCUs) in the reference summary, which correspond to an atomic unit of information. By design, the framework assumes there is more than one reference available for each document, and the frequency of SCU among all references acts as its weight (weights correspond to the level of importance, thus, the *pyramid*). In the next step, SCUs are extracted from the hypothesis and matched to the reference SCUs. The numerical score is computed by counting matching SCUs, giving more importance to the ones appearing in more references. While the process has clear advantage – SCU extraction and matching is far less subjective than direct scoring, allowing in principle a comparison between different documents and annotators – the annotation cost is high. Shapira et al. (2019) proposed the LitePyramid protocol, that simplifies the process by sampling SCUs from the union of references (without the inner-reference matching), and assigning binary labels (whether the SCU can be inferred from the hypothesis) to a sample of SCUs, without extracting SCUs from the hypothesis. The idea was recently re-visited by Liu et al. (2023b), who argue that the extraction should be done by experts, with the matching step benefiting more from the *scale* of evaluation. They also propose a modified formula for the final score that incorporates the length of reference and hypothesis, similar to the brevity penalty in BLEU (Papineni et al., 2002).
- **Utility-based** – An alternative approach to summary evaluation aims to measure the level of *utility* by measuring the performance degradation on an external task if the hypothesis is used in place of the reference. One such method is through the lenses of QA and was explored by, e.g., Liu and Lapata (2019a) or Li et al. (2020d). In their formulation, a number of human annotators are employed to create a set of questions (and answers) based on the reference summary. Next, the annotators were provided with the system summaries and were asked to answer the relevant questions,

without looking at the reference. Finally, a third round of annotations was conducted to mark the answers either as correct (score 1.0), partially correct (score 0.5), or incorrect (score 0.0). The final system score is the average over all of the questions and articles. While very costly, these kinds of annotations are more in line with practical applications where the ultimate goal is the satisfaction of the end user.

Instructions

In this task you will evaluate the quality of summaries written for a news article.
To correctly solve this task, follow these steps:

1. Carefully read the news article, be aware of the information it contains.
2. Read the proposed summaries A-F (6 in total).
3. Rate each summary on a scale from 1 (worst) to 5 (best) by its *relevance*, *consistency*, *fluency*, and *coherence*.

Definitions

Relevance:
The rating measures how well the summary captures the key points of the article.
Consider whether all and only the important aspects are contained in the summary.

Consistency:
The rating measures whether the facts in the summary are consistent with the facts in the original article.
Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

Fluency
This rating measures the quality of individual sentences, are they well-written and grammatically correct.
Consider the quality of individual sentences.

Coherence:
The rating measures the quality of all sentences collectively, to the fit together and sound naturally.
Consider the quality of the summary as a whole.

Article

\$(article)

Summaries

Summary A

\$(grounding)

Relevance	<input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input type="button" value="5"/>
Consistency	<input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input type="button" value="5"/>
Fluency	<input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input type="button" value="5"/>
Coherence	<input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="4"/> <input type="button" value="5"/>

Figure 3.4: Illustration of the evaluation framework that explores the Likert scale. Figure reprint from [Fabbri et al. \(2021\)](#)

In the following sections, we will provide an overview of the automatic metrics most commonly used to evaluate summarization systems. We will focus on metrics that aim to measure mostly the overall quality, noting the specialized lines of work on, e.g., factual consistency ([Kryscinski et al., 2020](#); [Xie et al., 2021](#); [Honovich et al., 2022](#); [Gao et al., 2023](#)).

3.1.2 String-based metrics

The ROUGE metric ([Lin, 2004](#)) measures the lexical similarity by counting the number of overlapping tokens (n-grams, word sequences, word pairs) between the

reference summary (summaries) and the one created by an automated system. It was inspired by the successful applications of BLEU and developed based on findings from the early DUC evaluations (Lin and Hovy, 2003). The original paper does not propose a single metric, but rather a family of metrics, that differ with the lexical units used to compute the overlap. The most commonly used variants are ROUGE-1 (overlap of unigrams), ROUGE-2 (overlap of bigrams), and ROUGE-L (based on the length of the Longest Common Subsequent). In contrast to the precision-based nature of BLEU, which explores the brevity penalty to punish too long translations, ROUGE takes a different approach. Instead, it is a common practice to report either both precision (fraction of the lexical units from the hypothesis present in the reference) and recall (fraction of the lexical units in the reference present in the hypothesis) or the combined F-measure (typically F_1 , i.e., the harmonic mean).

Despite its simplicity and a number of works criticizing ROUGE for low correlation with human judgments (see, e.g., Fabbri et al. (2021)), it is still often the single metric being reported for SOTA summarization systems (see, e.g., Touvron et al. (2023b)). Additionally, as recently highlighted by Grusky (2023), there is a lot of inconsistency when it comes to reporting ROUGE scores. A number of ROUGE implementations exist⁷ that differ by, e.g., the usage of stemmer or the approach to sentence splitting. What makes it even worse is that researchers fail to report not only the implementation they use but sometimes even the ROUGE variant, i.e., it is not clear whether the reported results are ROUGE-1 or ROUGE-L and whether it is precision, recall, or F-score. While other lexical-similarity-based metrics such as BLEU or METEOR (Banerjee and Lavie, 2005) are sometimes reported for summarization systems, their usage is rather marginal and ROUGE is by far the most prevailing one.

3.1.3 Embedding-based metrics

Building upon the observation that the exact match required by token-based metrics may be too strict – it does not consider partial matches nor word similarities – metrics that compute the similarity based on *token embedding* were proposed, with BERTScore (Zhang et al., 2020b) getting the most attention. Despite being proposed initially for MT and image captioning evaluation, it turned out to be an effective metric for summarization evaluation. BERTScore uses a pre-trained encoder – BERT (Devlin et al., 2019) in the original⁸ implementation, with RoBERTa-large (Liu et al., 2019) recommended at the time of writing – to obtain embeddings $H = (h_1, h_2, \dots, h_n)$ of the hypothesis and $R = (r_1, r_2, \dots, r_m)$ of the reference. Then, it computes precision (P) and recall (R) as follows:

$$P = \frac{1}{n} \sum_{i=1}^n \max_{j \in 1 \dots m} \langle h_i, r_j \rangle$$

$$R = \frac{1}{m} \sum_{j=1}^m \max_{i \in 1 \dots n} \langle h_i, r_j \rangle$$

⁷The original implementation of Lin (2004) known as ROUGE-1.5.5 was written in Perl, and to the best of our knowledge, currently is only available through third-party re-uploaders.

⁸https://github.com/Tiiiger/bert_score

with $\langle \cdot, \cdot \rangle$ indicating an inner product.

A different way of using pre-trained encoders for evaluation purposes was proposed by [Colombo et al. \(2022\)](#). The InfoLM metric they propose recursively masks each token position in both the hypothesis and the reference, remembering the discrete, per-token distributions. Next, the distributions are aggregated with a weighted average and compared with an information measure, such as the Kullback-Leibler divergence. It is clear that if $H = R$, then InfoLM will evaluate to 0. In the paper, the authors also show the contrary being true as well, i.e., if InfoLM evaluates to a substantial score, H , and R differ substantially, as each token h_i is unlikely given r_1^m . Compared to BERTScore, this approach does not require a calibration related to, e.g., the selection of a layer used to compute embeddings. A similar approach is proposed by MaskEval ([Liu et al., 2022c](#)), which measures the per-token contribution on the concatenation of source and hypothesis, allowing application in scenarios where reference summary is not available.

3.1.4 QA-based metrics

A line of research explored the QA paradigm as an indirect way of evaluating automatic summaries. The premise is as follows: “The automatic summary should suffice in practical applications as a substitute for a human-generated summary if the amount of information it carries – with the information amount judged by being able to answer similar questions if provided as a context – is close enough.”

[Eyal et al. \(2019\)](#) proposed the APES metric that used the reference summary to produce a fill-in-the-blank type of questions by finding all possible entities using a NER system. The APES score for a given summarization model is the percentage of questions that were answered correctly (using an automatic QA system), given the automatic summary as a context. [Scialom et al. \(2019\)](#) extended their work into unsupervised settings by generating questions from the source document. A follow-up work by [Durmus et al. \(2020\)](#) – FEQA metric – and [Wang et al. \(2020\)](#) – QAGS metric – automatically generates the natural language questions from the summary and/or document, no longer relying on fill-in-the-blank templates (see Figure 3.5). [Deutsch et al. \(2021a\)](#) and [Deutsch and Roth \(2022\)](#) made further improvements to the pipeline by taking a closer look at the answer comparison step and examining the impact of low-quality questions. [Manakul et al. \(2023\)](#) employs multiple-choice questions, and instead of comparing the most probable ones, compares the distributions over all possible choices.

Inspired by those works, in [Krubiński et al. \(2021a\)](#), we proposed the MTEQA metric that applies the same principles to MT evaluation. To compare our solution against other SOTA metrics, we participated ([Krubiński et al., 2021b](#)) in the WMT21 Metric Shared Task ([Freitag et al., 2021](#)), achieving the highest correlation with human annotators on the challenging English→Chinese test-set based on TED talks. A thorough summary of MTEQA is provided in Appendix A of this thesis.

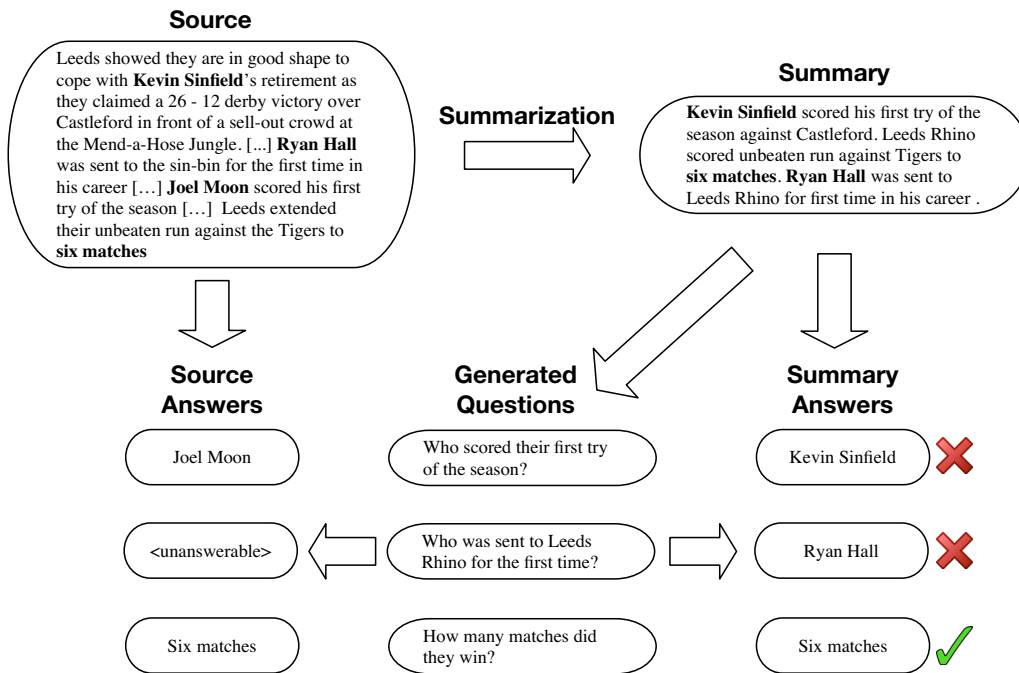


Figure 3.5: Overview of the QAGS metric. A set of questions is generated based on the summary. The questions are then answered using both the source article and the summary. Corresponding answers are compared using a similarity function and averaged across questions to produce the final QAGS score. Figure reprint from Wang et al. (2020)

3.1.5 LLM-based metrics

The GPT-3 paper (Brown et al., 2020) was one of the first to show that with enough training data and enough trainable parameters⁹, a rather simple autoregressive language model trained with the causal language modeling (next token prediction) task can be applied to solve a variety of tasks without any gradient updates or fine-tuning. Instead, it suffices to provide *a few* solved examples in the context, and the model should be able to deduct not only the task but also, hopefully, the correct answer. For example, given a context of `This is awesome!||Positive; This is bad!||Negative; Wow that movie was great!||Positive; What a horrible show!||` one can expect the implicit task of textual entailment to be guessed, and the correct label of `Negative` to be generated with future tokens. With the usage of instruction tuning (Wei et al., 2022a; Ouyang et al., 2022), which fine-tunes the model on explicit instructions followed by correct predictions, one can drop the requirement to provide examples, and simply prompt the model with commands. Those *emerging abilities* (Wei et al., 2022b) are still not fully understood, as it was shown (Min et al., 2022) that, e.g., swapping the labels in the context to random ones barely hurts the performance.

Considering the impact LLMs had on the whole field, it is not surprising that they have also been explored as an automatic tool to measure the quality of gen-

⁹There is no commonly accepted “size”, measured in the number of parameters, a model should have to be called “large”. We will follow the current trends and refer to the language models with few-shot or prompting capabilities as LLMs.

erated texts. Fu et al. (2023) proposed the GPTScore that assigns the cumulative log-probability of hypothesis h , given source S , aspect definition a (in the case of summarization, this would correspond to, e.g., fluency), and task description d as a context, i.e., $\log p(h_t | \mathbf{h}_{<t}, T(d, a, S), \Theta)$, with $T(\cdot)$ corresponding to the prompt template and Θ referring to the parameters of a particular model. After Kocmi and Federmann (2023) have shown that LLMs are capable of *directly* scoring a MT output with an appropriate prompt, e.g.,

```
Score the following translation from {source_lang} to {target_lang} with respect to
the human reference on a continuous scale from 0 to 100, where a score of zero means
"no meaning preserved" and score of one hundred means "perfect meaning and grammar".
```

```
{source_lang} source: "{source_seg}"
{target_lang} human reference: {reference_seg}
{target_lang} translation: "{target_seg}"
```

Score:

further works (Wang et al., 2023a) have shown that a similar approach also works for text summarization evaluation. Another usage of LLMs as an evaluation tool was proposed by Liu et al. (2023c), who, in the prompt, provide the source document with several candidate summaries and specify an instruction that asks the model to rank the summaries based on their quality.

For a detailed taxonomy of Natural Language Generation (NLG) evaluation with LLMs, we refer the reader to a survey by Li et al. (2024).

3.1.6 Trainable metrics

Before LLMs allowed a few-shot or even zero-shot scoring, a commonly explored approach consisted of using historical data of numerical human annotations to train a regression model to predict the numerical score.

From the modeling perspective, those metrics use a pre-trained encoder to encode source/reference/hypothesis – or just source/hypothesis, in the case of quality estimation – then combine those into a common representation that later is projected to a single vector. Finally, the vectorized representation is processed by a stack of fully connected layers, and the whole model, usually with part of the encoder frozen, is trained with a variant of MSE/ L_1 loss and the human assigned quality score as a target. Since the largest collection of such annotated model outputs is available from the WMT Metric Shared Task¹⁰, it is no surprise that the trainable metrics (also called estimator-based metrics) were mostly explored in the context of MT. Metrics such as BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), xCOMET (Guerreiro et al., 2023) or MetricX (Juraska et al., 2023) have consistently over the years achieved the highest correlations with human scores, making use of novel encoders and evolving data augmentation techniques.

Despite those metrics being explored in the context of other NLG tasks, not much research targeting specifically summarization was published. Sellam et al. (2020) proposed to use the automatic ROUGE scores to simulate human annotations during pre-training but reported the performance only on MT and data-to-text tasks from the WebNLG Challenge (Gardent et al., 2017).

Therefore, in Krubiński and Pecina (2022), we proposed to explore COMET for evaluating Text Summarization systems. We introduced a variant of the

¹⁰<https://github.com/google-research/mt-metrics-eval>

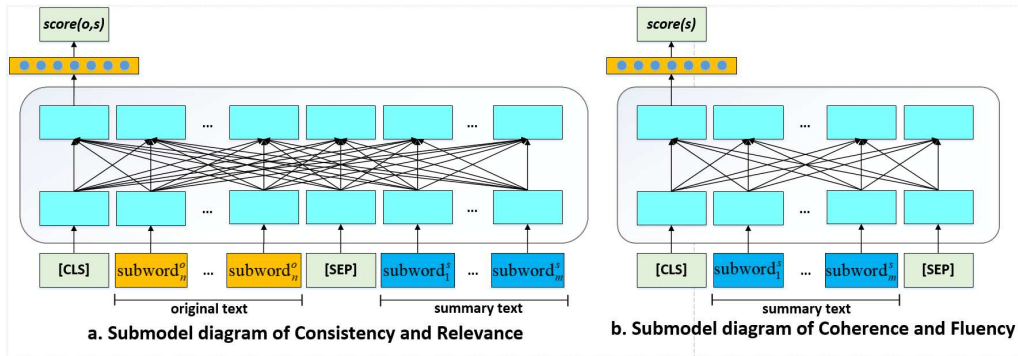


Figure 3.6: An overview of the trainable SummScore metric. Figure reprint from Lin et al. (2022)

model – COMES – trained on the annotated summarization outputs that used MT data for pre-training. In order to establish a proper benchmark, we examined the performance on several datasets with human judgments collected for different notions of summary quality, covering several domains and languages. We believe this was one of the first attempts to properly establish the role of trainable metrics in constrained settings, where the amount of human annotation is scarce, thus requiring a transfer learning from higher resource settings. Concurrently with our work, Lin et al. (2022) published the SummScore that explored only the SummEval (Fabbri et al., 2021) dataset, following a semi-supervised multi-round training regime. In their experiments, they train *separate* models for each evaluation dimension – Coherence, Consistency, Fluency, and Relevance – and do not explore transfer learning or task-specific pre-training. The authors use the source document and hypothesis as inputs when training with the Consistency and Relevance signals and just the hypothesis when training with the Coherence and Fluency signals (see Figure 3.6). Based on the rather limited scope of experiments and lack of official implementation, this approach did not catch a lot of attention¹¹.

¹¹As of May 2024, there is a single citation reported on Google Scholar.

3.1.7 COMES

This section is based on the FROM COMET TO COMES — CAN SUMMARY EVALUATION BENEFIT FROM TRANSLATION EVALUATION? (Krubiński and Pecina, 2022) article.

In this section, we will provide an overview of our work on the **COMES**¹² (Krubiński and Pecina, 2022) metric.

As briefly discussed in the previous section, one of the factors limiting the research on trainable summarization metrics is the amount of annotated data. At the time of writing the COMES paper, the largest datasets of source/reference/hypothesis triples with human annotations (Fabbri et al., 2021; Bhandari et al., 2020; Maynez et al., 2020) could be stretched up to thousands of instances, by, e.g., treating the same source/reference/hypothesis triple scored by three different annotators as separate instances. In reality, the data was even more limited, as the number of separate source articles that have annotated summaries is less than one thousand. On the contrary, the amount of annotated source/reference/hypothesis triples within the MT task was at that time close to 1M instances, thanks to the WMT conferences (Koehn et al., 2023, 2022; Barrault et al., 2021, 2020b). Building upon this, the question we asked is: *Can we use this resource to improve summary evaluation?* While the tasks of Machine Translation and Text Summarization are different, our research was built upon the belief that the problem of evaluating the quality of generated output is closely related.

To address this question, we examined the applicability of the COMET – a metric that is trained on the annotated MT data and capable of directly regressing a quality score – for summary evaluation. We proposed a variant of the model, COMES, that uses the annotated MT data for pre-training and is capable of predicting several aspects of summary quality simultaneously. In our experiments, we mostly explored the SummEval¹³ (Fabbri et al., 2021) dataset. It consists of 100 articles randomly sampled from the test split of the CNN/DailyMail corpus (Nallapati et al., 2016), each of them summarized by 17 systems. For each system output, the authors collected 3 expert judgments for several evaluation dimensions, i.e., *Coherence*, *Consistency*, *Fluency* and *Relevance* on a Likert scale of 1 to 5. In addition to the original reference, for each article, 10 alternative references were created by Kryscinski et al. (2020). As the COMET metric trained on the MT data outputs a single overall score, when reporting COMET performance, we compared this single overall score to all evaluation dimensions. To enable (semi-independently) predicting several aspects of summary quality at once, we proposed a modification that alters the number of outputs in the last feed-forward layer (see Figure 3.7). We experimented with both training from scratch (COMES) and pre-training on the annotated MT data by initializing the model weights from the COMET checkpoint (COMES_MT). In both scenarios, we examined the reference-less variant of the metric (COMES_QE and COMES_QE_MT, respectively).

¹²Crosslingual Optimized Metric for Evaluation of Summarization

¹³<https://github.com/Yale-LILY/SummEval>

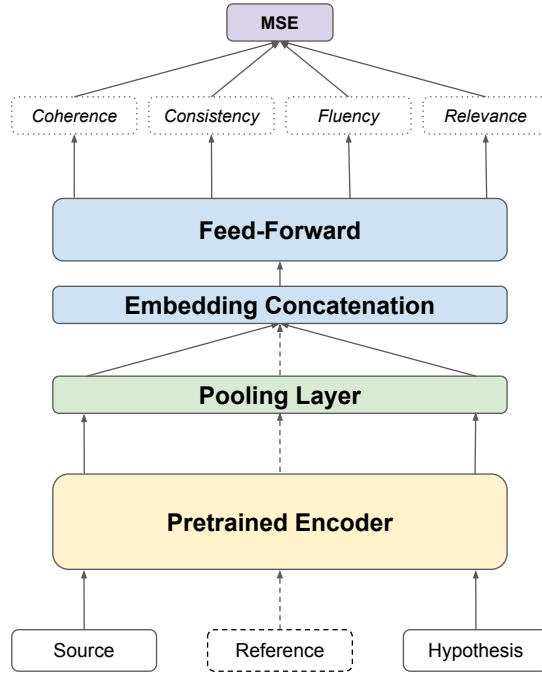


Figure 3.7: Estimator model architecture used in COMES. Source, reference, and hypothesis are all independently encoded with a pre-trained encoder. The pooling layer is used to create sentence embeddings from sequences of token embeddings. In the COMES variant, the last feed-forward layer has 4 outputs corresponding to different summary evaluation dimensions. Dashed lines are used to indicate the reference-less variant. Figure reprint from Krubiński and Pecina (2022).

At the time of working on this article, SummEval was the largest resource for summary evaluation. Thus, we wanted to use it both for training and evaluation. To achieve this, we relied on cross-validation. We split the data into 10 subsets of 10 articles each, using 80 articles for training, 10 for validation (early stopping), and evaluating the remaining 10. We trained 10 models, used each of them to score 10% of the available (unseen) data, and merged the results. That way we could directly compare to other metrics that report correlation on the whole SummEval dataset. During training, we used each reference and each expert annotation¹⁴ to create more training instances (80 articles \times 11 references \times 17 models \times 3 annotations = 44,880 instances). During evaluation, we handled multiple references by scoring each reference independently and taking the maximum score.

Our baseline experiments showed that scoring system outputs with both out-of-the-box variants (COMET and COMET_QE) resulted in the highest correlation coefficients along all metrics analyzed by Fabbri et al. (2021) for *Coherence* and *Relevance* dimensions (see Table 3.1). The reference-less variant (COMET_QE) had a much higher correlation with the *Consistency* dimension. Both COMES and COMES_QE variants performed similarly, achieving higher correlations than both COMET and COMET_QE. However, the effect of pre-training (COMES_MT) was ambiguous – on average, it did not help, but the

¹⁴We have tried averaging human ratings during training, the results were comparable but slightly worse.

Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-3 F1	0.2206	0.7059	0.5092	0.3529
ROUGE-4 F1	0.3088	0.5882	0.5535	0.4118
BERTScore F1	0.2059	0.0441	0.2435	0.4265
CHRF	0.3971	0.5294	0.4649	0.5882
METEOR	0.2353	0.6324	0.6126	0.4265
COMET	0.5735	0.2353	0.5240	0.6765
COMES	0.6912	0.7206	0.5830	0.7206
COMES_MT	0.6471	0.4412	0.6273	0.7206
COMET_QE	0.4118	0.7206	0.7011	0.5441
COMES_QE	0.6618	0.7647	0.6126	0.7059
COMES_MT_QE	0.6912	0.4853	0.6126	0.6912

Table 3.1: System-level Kendall’s τ correlations with (average) expert annotations for four evaluation dimensions annotated in the SummEval dataset. The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).

main cause was the poor performance on predicting the *Consistency* dimension.

To get a better understanding of the metric performance, we applied it to several other annotated summarization datasets¹⁵. Since we trained 10 instances for each variant of the COMES models, evaluating with each of them allowed us to estimate the confidence intervals directly, not having to rely on, e.g., bootstrapping (Deutsch et al., 2021b).

To examine the performance on non-matching evaluation dimensions, we reported results on data¹⁶ from the same domain – a subset of the CNN/DailyMail corpus. Bhandari et al. (2020) produced the numerical gold-standard scores by rating a system output based on a number of Semantic Content Units (SCUs) that can be inferred from it. LitePyramid (Shapira et al., 2019) method was used to obtain SCUs from reference summaries. On this dataset (see Table B.1), the reference-less COMET_QE outperformed any other variant, almost doubling the correlation of COMET. The *Consistency* head of COMES_QE came in second. We observed the best correlation to be obtained by the recall variant of ROUGE, which can be explained by the recall-based nature of annotations. In an independent work¹⁷, Stiennon et al. (2020) annotated a different subset of the CNN/DailyMail corpus by rating system outputs for *Accuracy*, *Coherence*, *Coverage* and *Overall Quality*. On this dataset (see Table B.2), the reference-less variant COMET_QE performed best, obtaining almost a perfect correlation with the *Overall* dimension. This was by far a better result than any traditional metric considered. COMES trained from scratch outperformed the pre-trained variant COMES_MT, which may indicate overfitting to the SummEval annotations. Surprisingly, the highest correlation with the *Coherence* dimension (present in the SummEval annotations used for training) was not obtained by the *Coherence* head of COMES, but the *Relevance* one. The pre-trained variant (COMES_MT) did

¹⁵For clarity, the tables are available in Appendix B.1.

¹⁶<https://github.com/neulab/REALSumm>

¹⁷<https://github.com/openai/summarize-from-feedback>

not display this unexpected behavior. However, the correlation was marginally lower compared to the variant trained from scratch (COMES). To validate the performance on a different domain, we evaluated (see Table B.3) on the subset of the TL;DR corpus (Völske et al., 2017) annotated in a similar manner by Stiennon et al. (2020). On this dataset, COMET achieved the top correlation, with COMES clearly lagging behind in performance compared to the pre-trained COMES_MT variant.

One of the strengths of the COMET metric is its multilingualism – the model has seen over 30 language pairs during training. To assess its quality as a summary evaluation tool for non-English data, we evaluated it on the Multi_SummEval dataset (Koto et al., 2021). With only two system outputs annotated (along the *Focus* and *Coverage* dimensions), the size of the resource is not sufficient for reporting system-level correlations. Thus, we examined only the summary-level (segment-level) correlations. For a fair comparison, we wanted to train the COMES model variant using multilingual data. Due to the lack of sufficient resources, we fell back to using automatic MT to translate the English data. This approach has proven successful for, e.g., Question Answering (Lewis et al., 2020b; Macková and Straka, 2020). As we limited our analysis to the subset of languages from Multi_SummEval that originates from the MLSUM (Scialom et al., 2020) corpus, we have translated SummEval into German, French, Russian, Turkish and Spanish using the uni-directional models provided by the Helsinki-NLP group (Tiedemann, 2020) and used the data (together with the original, English SummEval) to train a multilingual COMES model (COMES_MT_ML). In the summary-level evaluation (see Table B.4), the original COMET metric was superior to any other variant considered, clearly outperforming the reference-less variant COMET_QE. Surprisingly, both the COMES_MT and the COMES variants performed better than the multilingual COMES_MT_ML variant. This is in line with the findings by Braun et al. (2022), which indicate that summary evaluations do not survive translation. On this dataset, even the best-performing COMET was still inferior to both ROUGE and BERTScore. Considering, however, the relatively small size of the dataset (270 instances per language, outputs from two systems), we considered the question about COMET/COMES applicability to multilingual evaluation to still be an open one.

As a part of ablation studies, we challenged the requirement of the cross-validation approach to the SummEval evaluation. We were able to show that the model trained and evaluated on the whole data was able to achieve almost the perfect correlation, indicating a strong over-fitting. Additionally, we approached a potential issue caused by the factor distinguishing MT from summarization. Namely, the average length of the input. While a typical segment in the MT test-sets consists of one or two sentences, the typical source document in summarization test-sets can be five or even ten times longer. By examining the lengths of the tokenized documents from the SummEval dataset, we estimated that only roughly 50% of them fit completely within the model limit of 512 tokens¹⁸. However, we computed that, on average, 92% of input tokens are consumed, as an average input document length in tokens equals 502. We were not able to propose any additional experiments that could numerically estimate the potential gap in

¹⁸In all of our experiments, we have explored the large variant of the XLM-RoBERTa (Conneau et al., 2020) encoder.

performance compared to a variant that could consume 100% of input tokens.

By the end of the paper, we concluded that the gap between the off-the-shelf COMET and the fine-tuned COMES is, in our opinion, not significant enough to justify relying on COMES, recommending using either the COMET or COMET_QE to measure the overall quality. This recommendation was taken into account by (Lozano et al., 2024), who, in their research, used the COMET metric to evaluate the Retrieval-Augmented LLM system.

3.2 Visual Output

As discussed in Chapter 2, our focus is the supervised formulation of MSMO with a single image in the output. In those settings, our input is either a set of images $I = (i_1, \dots, i_n)$ or a video $V = (v_1, \dots, v_m)$ represented by the sequence of frames. For simplicity, we assume that the pre-processing step of frame sampling was already performed, and all of the m frames are seen by the model. In both cases, we assume there is a gold-standard pictorial reference r and that the final goal of an automatic system is picking a single image/frame from the input.

The crucial difference between the image-based and video-based formulation was discussed in Section 2.1. In the image-based formulation, we assume that the reference image belongs to the input images, i.e., $r \in I$, or $\exists j : i_j = r$. In the video-based formulation, we use a metric m to compute similarity $m(v_j, r)$ between each frame v_j and reference image r . In practical applications, the most common metric m is the cosine similarity¹⁹ computed between the vectors corresponding to the (target) image and the (input) frame features. Namely, we use a pre-trained feature extractor (see Section 1.2.1) to embed the image i into a vector $t = t(i) \in \mathbb{R}^d$. Finally, we either treat the frame v_k most similar to the reference image, i.e., $k = \arg \max_j m(v_j, r)$ as the gold standard, and thus converge towards the image-based formulation, or directly report the similarity $m(v_{pred}, r)$ between the reference image and the frame v_{pred} picked by the model. An alternative approach would be to restrain from using the arg max formulation and instead consider all frames that are *similar enough* to the reference as positive labels. This could be achieved by choosing a particular threshold τ , and assuming that every frame v_k with $m(v_k, r) \geq \tau$ has a positive label. The concept of positive/negative labels is also relevant to the image-based formulation, as some of the human evaluations conducted in previous research (see, e.g., Zhu et al. (2018)) mark more than one input image as “relevant”.

The second crucial distinction comes from the modeling design. We either work with models that directly pick a particular image/frame by, e.g., generating the index (see Section 5.2) or with models that assign a score s to every input image/frame, i.e., $s(v_j) = s(v_j, \Theta)$, with Θ indicating model parameters. If we look only at the index of the top-scoring sample, then we fall back to the previously discussed case.

If the model picks a single image/frame, then the Accuracy metric (a proportion of predictions where the reference image is currently retrieved by the model) is a natural choice for the image-based formulation, with the raw similarity score $m(v_{pred}, r)$ filling the same role for the video-based variant.

If we consider all of the scores, i.e., a ranking of all input samples, then the task is typically looked at from the perspective of Information Retrieval. This approach is better suited for formulations that allow more than one positive label among the input images/frames. In that case, the typically reported metrics are

¹⁹From the mathematical point of view, cosine similarity between any two vectors can range from -1 to 1 . The experience shows that for vectors corresponding to features extracted with the same neural network, the value is usually positive and thus can be treated as a measure of similarity.

Precision@k (a proportion of top-k predictions that are relevant), Recall@k (a proportion of relevant samples within the top-k predictions), and Mean Average Precision (computed based on values of Precision@k, averaged over several values of k).

When interpreting and comparing those “visual” metrics between different works and different datasets, it is important to remember that certain pre-processing or implementation-related choices may heavily influence the final scores. For example, if two separate feature encoders are employed, the distribution of cosine similarity scores may drastically change. Additionally, the (average) amount of input images/frames affects the classical baselines. Considering the difficulty of the task, this is of crucial importance, as even the best-scoring models are not achieving much higher scores than classical baselines (see Section 2.2 and Chapter 5). While a simple Accuracy is free of such issues, besides the top-1 output, it does not tell us much about the model’s performance. One can imagine a practical application that would require a summary consisting of two or three images, and in such a case, the whole distribution of scores (the ranking) becomes a critical factor.

3.3 Multimodal Output

Parts of this section are based on the MLASK: MULTIMODAL SUMMARIZATION OF VIDEO-BASED NEWS ARTICLES (Krubiński and Pecina, 2023) article.

Within the discussed formulation of supervised *MSMO* (see Chapter 2), both the model output – (i_{pred}, txt_{pred}) – and the reference – (i_{ref}, txt_{ref}) – consist of a <image, text> pair. We believe that in order to fully grasp the quality of the multimodal output – both the text and image are supposed to be presented to the final user as a whole – problem-specific, multimodal metrics should be employed. However, almost all works report two distinct sets of metrics – the output text is compared to the reference text, and the output image is compared to the reference image. The main reason for that is simply the lack of well-established multimodal metrics. Unfortunately, this issue can not be solved easily – in order to design (and validate) automatic metrics, a collection of annotated model outputs is required. Taking aside the practical problems related to, e.g., the cost of a large-scale annotation, the field lacks every component required for such a process:

- there are no standard evaluation frameworks (see Section 3.1.1 for the discussion on evaluating text-only output);
- a majority of works operate on internal datasets that for various reasons, are not shared publicly (see Chapter 4);
- the public code-bases and models are mostly tailored for specific datasets/features and can not be easily evaluated with even a slightly different setup.

Still, a limited number of solutions (see below) have been proposed both in terms of evaluation design and automatic metrics.

Human Evaluation

The first work (Zhu et al., 2018) to introduce the *MSMO* task already noticed the challenging nature of multimodal evaluation. In their image-based formulation, there is no gold-standard reference picture, but instead, the authors employed graduate students to select up to three *relevant* images per article. To collect human annotations for the pictorial summaries (model output), annotators were requested to judge the relevance of the output in the context of a gold-standard textual summary and the subset of images marked as relevant in the previous step. In total, 600 samples were annotated on the Likert scale of 1 to 5, with each sample scored by two people. Interestingly, to the best of our knowledge, no other work on *MSMO* conducted an evaluation at a similar scale. Even the first work (Li et al., 2020d) to introduce the video-based *VMSMO* formulation did not evaluate the pictorial summaries, collecting human judgments only for the textual output.

For the purpose of the experiments reported in the MLASK (Krubiński and Pecina, 2023) paper, we designed our own cross-modal evaluation framework, better suited for datasets with a single, gold-standard pictorial summary. Since the input images are actually frames sampled from a video, besides the *relevance*,

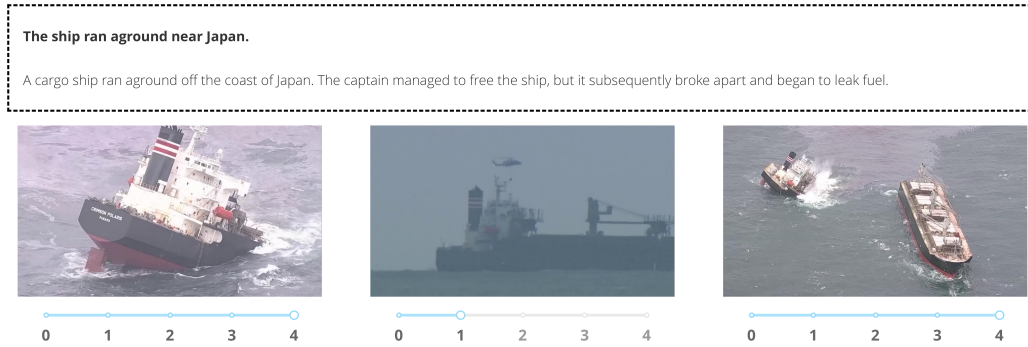


Figure 3.8: Screenshot of the annotation tool used to collect human judgments about the quality and usefulness of selected cover frame. For convenience, we translated all [text](#) into English. Figure reprint from [Krubiński and Pecina \(2023\)](#).

we wanted the annotators to also look at the quality. It may happen that a particular frame sampled from the video is blurred, as it may resemble e.g., a fading away end of a scene.

To evaluate the quality of cover frame selection, we asked human annotators to judge the quality and usefulness of an image as a pictorial summary of the article. Altogether, 18 human annotators participated. All were adult, native Czech speakers (the textual part of the MLASK dataset is in the Czech language) who read online news magazines daily. Figure 3.8 displays a screenshot of the annotation tool. For each instance, the annotators were asked to rate 3 images on a Likert scale of 0 to 4 (the higher, the better) in the context of the article’s title and the reference summary.

The suggested interpretation of the scale levels was:

- 0: The picture is not relevant at all or very marginally (technical quality is not important).
- 1: The image is partly relevant (there is a certain connection between what it captures and the content of the text), but technically imperfect (e.g., blurred, cropped inappropriately, taken from an inappropriate angle or at an inappropriate moment).
- 2: The image is partly relevant (there is a certain connection between what it captures and the text content) and of good technical quality.
- 3: The picture is very relevant but technically imperfect (e.g., blurred, cropped inappropriately, taken from an inappropriate angle, or at an inappropriate moment).
- 4: The picture is both very relevant and of good technical quality. It is a suitable cover picture.

Compared to the annotation design by [Zhu et al. \(2018\)](#), we did not provide the reference image as a context, instead including it as one of the options. Such a premise allowed us to get closer to what we consider a real-life settings, i.e., a situation where the model output is used instead of the reference. To set up the annotation, we randomly chose 300 instances from the MLASK test-set and split them into 10 batches of 30 instances each. We used the first batch to measure the inter-annotator agreement, asking each annotator to score all the instances in the control batch plus at least one more batch. For each instance, four images were considered for annotation: the reference picture, a random frame from the video,

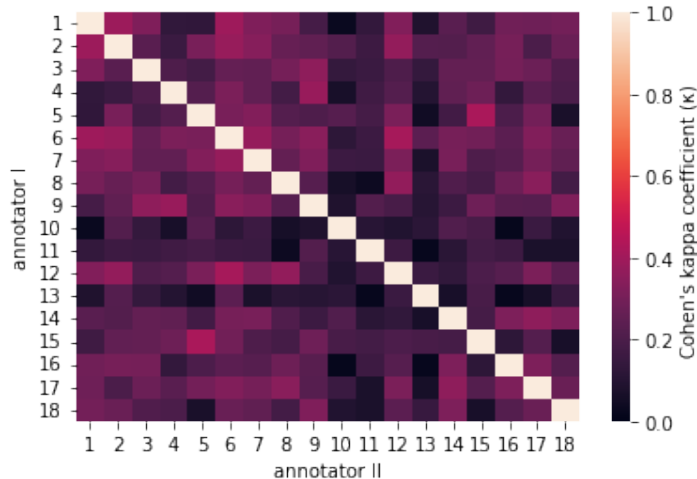


Figure 3.9: Values of Cohen’s κ used to measure the inter-annotator agreement on the control batch. Figure reprint from Krubiński and Pecina (2023).

and the outputs of two test models. In the control batch, we always included the reference picture, hiding the output from one of the methods in 33% of the cases. In the other batches, we displayed three out of the four images selected randomly. The reason for such a design is purely practical – during our early experiments, we noticed that the annotation process was much smoother if all of the images were presented side by side. When viewed on an average computer screen, four images presented at once seemed too clunky and hard to read. To avoid positional bias, we shuffled the images each time before showing them to the annotator. On average, we collected 2.5 annotations for each image.

Figure 3.9 displays the inter-annotator agreement measured on the control batch in the form of a heat map. The agreement is measured with Cohen’s κ coefficient (Cohen, 1960), with the average value of 0.217 indicating a “fair” agreement. One can notice that three annotators (10, 11, and 13) have a lower average agreement (average below 0.2, indicated by the dark stripes). As a precaution, we decided to exclude their annotations from further analysis. By doing so, the average value of Cohen’s κ increased to 0.26, and the average number of annotations decreased to 2.2.

The results related to the model performance are discussed in Section 5.1.3.

Automatic metrics

As announced in the previous section, due to the lack of a large-scale, publicly available collection of annotated model outputs, very limited research approached the creation/design of automatic metrics tailored for multimodal output.

One such metric, MMAE, was proposed by Zhu et al. (2018). Using the 600 annotations they collected, 450 were set aside to train the MMAE metric, and the remaining 150 were used to evaluate its performance. The MMAE metric consists of three components: text-to-text similarity measured with ROUGE-L, image-to-image similarity measured with Image Precision (based on the retrieval formulation, recall is not relevant as there is a single image in the prediction), and the cross-modal correspondence MAX_{sim} (an image-to-text retrieval model

is employed to calculate the similarity between i_{pred} and every sentence in txt_{ref} , with a max-based approach to aggregation). The MMAE metric is the linear combination of those three, with the coefficients learned by either a Linear Regression, Logistic Regression, or a simple (shallow) feed-forward network – with the best results achieved by Linear Regression. Evaluation on the test-split of the annotated outputs shows that the combined metric achieved a higher correlation than any of the components by itself. In their follow-up work, [Zhu et al. \(2020\)](#) extend the MMAE metric (MMAE++) by incorporating a fourth component based on the cross-modal, segment-level retrieval system into the regression setup. Namely, the authors sample image-captions pairs to create a collection of matching multimodal segments. Given segments $m_a = (i_a, txt_a)$ and $m_b = (i_b, txt_b)$, a pair of *matching* segments (m_a^*, m_b^*) is created by swapping text and image, i.e., $m_a^* = (i_a, txt_b)$ and $m_b^* = (i_b, txt_a)$. Based on such pairs, a retrieval system is trained to retrieve pairs of matching segments. Finally, this retrieval system is used to compute a matching score between (i_{pred}, txt_{pred}) and (i_{ref}, txt_{ref}) , with a max-based approach to multiple i_{ref} images. Using the test-set of 150 annotations, the MMAE++ with four components scores better than the simpler MMAE version.

Both the MMAE and MMAE++ metrics not only achieve high correlations with human annotations but also incorporate every aspect that we deem important for evaluating the quality of multimodal output – text-to-text correspondence, image-to-image correspondence, and cross-modal dependencies. However, they are unfortunately not applicable from a practical point of view. The coefficients are dependent on a particular dataset, a collection of particular model outputs, and a particular group of human annotators. On top of that, even if one collects a significant enough collection of annotated model outputs to justify “generic” coefficients, they would be tied to specific models used in, e.g., the retrieval systems (one of MMAE/MMAE++ components). Switching to a better-performing retrieval model would either require some form of convoluted score rescaling or another round of coefficient calibration. Therefore, we treat both the MMAE and MMAE++ metrics as great proof-of-concept initiatives but restrain from applying them to our own research. Overall, those issues are yet another hint at how difficult and complicated the [MSMO](#) task is and that a proper, task-specific evaluation is key to the further development of this field.

It is also worth mentioning that some work was done on evaluating Multimodal Summarization with text-only output, specifically within the “Text + Image \rightarrow Text” formulation. [Wan and Bansal \(2022\)](#) proposed the CLIPBERTScore metric, a weighted, linear combination of CLIPScore ([Hessel et al., 2021](#)) and BERTScore ([Zhang et al., 2020b](#)). CLIPScore is based on the CLIP image-text retrieval model and was proposed as a metric to detect hallucinations in image captions. It is applied to the input image and the textual summary generated by the model. A multi-image input is also considered, with an additional step that averages the score over all input images. BERTScore (see Section 3.1.3) is computed between the (textual) hypothesis and the (textual) input document. To validate the proposed metrics, the authors annotate summaries of 50 articles from the WikiHow dataset ([Koupae and Wang, 2018](#)), enhanced with input images by [Yang et al. \(2021\)](#). Their validation shows that the proposed CLIPBERTScore

outperforms any other metric considered. Interestingly, a simple grid-based tuning of the coefficient used in the linear combination (no intercept, coefficients sum to 1.0) performs on par with a calibration based on Linear Regression or simple feed-forward network, both of which were explored by [Zhu et al. \(2018\)](#) to calibrate MMAE. Compared to MMAE, CLIPBERTScore is more convenient to apply – it is based on two well-established models and includes only a single parameter. We explored it in our experiments on unified, multi-task, multi-modal summarization (see Section 5.2), adapting it to multimodal output by switching the input image with the predicted one during the CLIPScore computation.

[Jing et al. \(2024\)](#) build upon the CLIPBERTScore work to introduce the FALLACIOUS framework for both reference-based and reference-free evaluation. FALLACIOUS is a QA-based metric (see Section 3.1.4) that, in the first step, generates binary (yes/no) questions based on the textual hypothesis. The numerical score is produced by counting the percentage of questions that are correctly answerable based on either the input document or the input image. Compared to CLIPBERTScore, this approach does not rely on any coefficients that require calibration. It is, however, more costly (due to the QA/QG steps) and prone to the same problems as QA-(text-)based evaluation, e.g., a lack of sensitivity to abstract terms.

4. Datasets

4.1 Overview

In Chapter 2, we discussed the task of Multimodal Summarization, focusing mostly on taxonomy and problem variations approached previously. In this chapter, our goal is to provide an overview of the datasets explored in previous research and to introduce our contributions – a curation of the MLASK dataset (see Section 4.2) and the extension of the existing M3LS dataset (see Section 4.3). We will limit ourselves to the datasets that are publicly available for download, with a single coherent document in the input. As discussed before, Multimodal Summarization was mostly explored in the news domain, and the datasets were often curated by web scraping a particular website. Therefore, publicly sharing the datasets curated by researchers is not always possible due to licensing and intellectual property rights. Works that explored the task in, e.g., the e-commerce domain, were often published by authors affiliated with industrial research companies. Thus, the artifacts – code, models, and datasets – are kept private and not meant to be shared with the public at all. On top of that, due to the multimodal nature of the datasets – especially the ones containing videos – the size of the datasets may reach hundreds of gigabytes and require access to a specialized hosting platform, which may be a limiting difficulty for some research groups.

As a consequence, only a limited number of datasets are freely available to download. The ones that are published fall into one of the three data categories:

1. **Raw data** – datasets in this category can be downloaded as a whole, in the original, *raw* form, i.e., videos as `.mp4` and images as `.jpg` or `.png`. While this form enables a lot of freedom in future works, it may require substantial computational power for storing and processing.
2. **Set of pre-computed features** – works in this category do not share videos/images directly but rather publish frame/image-level features extracted with a particular feature extractor (see Chapter 1). While this kind of data enables relatively quick experiments and lowers the entry barrier, it forces the subsequent works to follow the same pre-processing (e.g., frame sampling or image cropping) and use the same feature extractors.
3. **Instructions to re-create the datasets** – a third category consists of works that do not publish the data at all but rather publish their setup – usually code and a collection of URLs – that can be used by others to gather the data. This approach is, however, often flawed. At first, a notable amount of time and computational resources are required to reconstruct the dataset. Secondly, most of the content available on the internet is not static, and URLs may become invalid as providers update their websites¹.

A list of publicly available datasets is presented in Table 4.1, with the corresponding sources provided in Table 4.2 .

¹In our experiments, we tried to reconstruct the video-based dataset introduced by Li et al. (2020d), roughly six months after the relevant paper was published. Out of the 184,920 articles collected in the original data, almost 93% of the corresponding URLs were no longer active.

Dataset	#Documents	Input Modalities	Output Modalities	Data category	Language
MMSS (Li et al., 2018)	66,000	T, I	T	1)	eng
How2 (Sanabria et al., 2018)	72,983	T, V	T	2), 3)*	eng, por*
MSMO (Zhu et al., 2018)	314,581	T, I	T, I*	1)	eng
VMSMO (Li et al., 2020d)	184,920	T, V	T, I	3)	zho
MM-AVS (Fu et al., 2021)	2,173	T, V, I	T	1), 3)*	eng
HCSCS-MS (Zhang et al., 2022a)	62,880	T, I	T, I	2)	zho
MLASK (Krubiński and Pecina, 2023)	41,243	T, V	T, I	1)	ces
M3LS (Verma et al., 2023)	1,100,000	T, I	T	1), 2)*	<i>various</i>
M ³ Sum (Liang et al., 2023a)	1,078,215	T, I	T	3)	<i>various</i>
mRedditSum (Overbay et al., 2023)	3,030	T, I	T	3)	eng
MultiSum (Qiu et al., 2024)	5,100	T, V	T, I	1)*, 2), 3)	eng

Table 4.1: Overview of the publicly available datasets explored for Multimodal Summarization. “T” refers to the textual modality, “V” to the video modality, and “I” to the image modality. In the “Language” column, we provide the three-letter [ISO 639-2](#) code. The * symbol indicates a partial match, e.g., only a part of the dataset (usually the test-split) is annotated, or only part of the data (images, but not videos) is released.

Dataset	URL
MMSS (Li et al., 2018)	https://github.com/ZNLP/ZNLP-Dataset/
How2 (Sanabria et al., 2018)	https://github.com/srvk/how2-dataset/
MSMO (Zhu et al., 2018)	https://github.com/ZNLP/ZNLP-Dataset/
VMSMO (Li et al., 2020d)	https://github.com/iriscxy/VMSMO/
MM-AVS (Fu et al., 2021)	https://github.com/xiyan524/MM-AVS/
HCSCS-MS (Zhang et al., 2022a)	https://github.com/LitianD/HCSCS-MSDataset/
MLASK (Krubiński and Pecina, 2023)	http://hdl.handle.net/11234/1-5135
M3LS (Verma et al., 2023)	https://github.com/Raghvendra-14/M3LS/
M ³ Sum (Liang et al., 2023a)	https://github.com/XL2248/D2TV/
mRedditSum (Overbay et al., 2023)	https://github.com/Koverbay/mredditsum/
MultiSum (Qiu et al., 2024)	https://mmsum-dataset.github.io/

Table 4.2: Overview of the publicly available datasets explored for Multimodal Summarization, with the corresponding URLs.

4.2 MLASK

This section is based on the MLASK: MULTIMODAL SUMMARIZATION OF VIDEO-BASED NEWS ARTICLES (Krubíński and Pecina, 2023) article.

In the early stages (first quarter of 2021) of our experiments, no video-based dataset with reference pictorial summaries was available publicly. Therefore, we focused our attention on curating a dataset that would enable our experiments and one that could be released to the public to facilitate further research. With that goal in mind, we decided to curate the dataset based on two news websites publishing in the Czech language: <https://novinky.cz> and <https://seznamzpravy.cz>. They both publish new articles daily and give access to an extensive archive of articles from previous years. Instead of scraping the web pages, we collected the documents via APIs, limiting our analysis to articles containing videos. By matching the API fields (see Figure 4.1) with the structure of the website, we were able to identify fields corresponding to: the article’s text, the article’s abstract, the article’s title, the article’s publication date, a .mp4 video that accompanied text, and a single image (cover picture) that visually represents the whole article (see Figure 4.2).

Having collected the articles, we processed them by filtering out documents that we had identified as invalid or of low quality. The following documents were dropped:

- with videos longer than 5 minutes;
- with full text shorter than 50 words or longer than 2,000 words;
- with abstract shorter than 10 words or longer than 80 words;
- with title shorter than 2 words;
- with either the full text or abstract identified as non-Czech by the `langid`² language-identifier.

The thresholds were chosen manually after exploring the distribution of texts.

In total, the collected dataset contains 41,243 instances, all including the article’s text, title, abstract, video, and cover picture. The quantitative statistics of the data are displayed in Table 4.3. The oldest article was published on the 22nd of September 2016, and the newest one on the 4th of February 2022. All of the videos are re-sampled to 25 `fps` and resized to the same resolution of 1280×720 p, and the average video duration is 85.58 seconds. All of the images are in the `.jpeg` format and are published with their original pixel size. The smallest image has a resolution (width \times height) of 199×229 , and the largest is 9000×6000 , with 40% of images having a resolution of 800×450 and 35% of 1920×1080 . We do not differentiate between the article’s origin, i.e., we treat both source websites equally by not labeling the documents with their origin. We named the dataset **MLASK**, which stands for **M**ultimoda**L** **A**rticle **S**ummarization **K**it.

By getting in contact with the media company owning the news websites that we scraped and receiving the necessary approvals, we were able to publish

²<https://github.com/saffsd/langid.py>

```

1  {
2    "_cls": "ArticlePublished",
3    "_created": "2022-09-03 08:05:12",
4    "_id": "6313849b63f3a49f51dab2da",
5    "adKeywords": [],
6    "authors": [
7      "5a9d100265375b136a38de6b"
8    ],
9    "content": [
10     {
11       "_cls": "ParagraphMolecule",
12       "component": "molecule.paragraph.Paragraph",
13       "componentId": "631384932b38ee436f9e76c2",
14       "properties": {
15         "_cls": "_ParagraphMoleculeProperties",
16         "entityRanges": [],
17         "inlineStyleRanges": [],
18         "text": "Dnes ve 20:17 SELČ se mělo na dvě
19           ↪ hodiny otevřít okno pro start rakety SLS
20           ↪ s modulem Orion k Měsíci. NASA už
21           ↪ předtím na svých stránkách uvedla, že má
22           ↪ potíže s únikem paliva a o několik hodin
23           ↪ později start zcela zrušila."
24       },
25       ...
26     },
27     {
28       "_cls": "ParagraphMolecule",
29       "component": "molecule.paragraph.Paragraph",
30       "componentId": "631384932b38ee436f9e76c3",
31       ...
32     },
33     ...
34   ]
35   "title": "NASA podruhé zrušila start rakety k měsíci.
36     ↪ Podívejte se, jak měla letět",
37   "uid": 213174,
38   "videoportalTags": [
39     "6311acf5e62d2906d836e71d"
40   ]
41   ...
42 }

```

Figure 4.1: An example ([Seznam Zprávy API](#)) of an output from the API call that we used to collect the documents. For clarity, only a subset of retrieved fields is presented.

Video ■

Article ■

Title ■

Summary ■

Image Summary ■

Zprávy » Tech » Technologie » NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měl...

NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měla letět

JAN MAREK

Podívejte se, jak má k Měsíci letět nejsilnější a nejdražší raketa n...
Článek

05:06

Nejsilnější raketa Země dnes letí k Měsíci

IMÉNO MODULU: SLS (Space Launch System)

VÝŠKA: 98,1 m

KAPACITA PŘI MĚKLAD K MĚSÍCI: 26,9 t

Podívejte se, jak má k Měsíci letět nejsilnější a nejdražší raketa na světě. (Video: Jan Marek)

3. 9. 17:54

Americký Národní úřad pro letectví a vesmír (NASA) dnes podruhé v tomto týdnu kvůli technickým potížím zrušil plánovaný start rakety SLS s modulem Orion k Měsíci. Vesmírná agentura o tom informuje na svých internetových stránkách.

Dnes ve 20:17 SELČ se mělo na dvě hodiny otevřít okno pro start rakety SLS s modulem Orion k Měsíci. NASA už předtím na svých stránkách uvedla, že má potíže s únikem paliva a o několik hodin později start zcela zrušila.

Šlo během jednoho týdne o už druhý pokus o vzlet vesmírného korábu s největším tahem, ale i cenou na světě. Jeho vývoj se totiž dost prodloužil i kvůli problému s motory, které přerušily start k Měsíci i toto pondělí. Brzy má přitom letět i s lidmi.

NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měla letět
3. 9. 17:54 - JAN MAREK

Figure 4.2: An annotated screenshot representing one of the articles (seznamzpravy.cz) collected in our experiments. The <image, text> pair in the bottom left corner corresponds to the thumbnail representing the article on the news provider’s main web page, with the “text” field given by the article’s title.

	Mean	Q_1	Median	Q_3
Title	11 ± 2.8	9	11	13
Abstract	33 ± 13.9	22	32	43
Article	277 ± 191.7	154	231	343

Table 4.3: Quantitative statistics of the lengths of titles, abstracts, and full texts (measured in the number of tokens) for the MLASK dataset. Q_1 and Q_3 denote the first and the third quartile, respectively. Table reprint from Krubiński and Pecina (2023).

the collected dataset via the LINDAT/CLARIAH-CZ data repository³, with an appropriate license⁴, freely permitting a usage in academic research. Separately, we also published the code⁵ that enables re-computing the numerical features explored in our experiments.

³<http://hdl.handle.net/11234/1-5135>

⁴<https://lindat.mff.cuni.cz/repository/xmlui/page/szn-dataset-licence>

⁵<https://github.com/ufal/MLASK>

4.3 Extension of the M3LS dataset

This section is based on the TOWARDS UNIFIED UNI- AND MULTI-MODAL NEWS HEADLINE GENERATION (Krubiński and Pecina, 2024) article.

When discussing the formulation of Multimodal Summarization with Unimodal Output that consumes a textual document and collection of images to generate a textual summary, we introduced the M3LS dataset curated by Verma et al. (2023). At the time of starting the work on the multi-task, multi-modal summarization (first quarter of 2023, see Section 5.2), it was the largest (376,367 instances) resource of such kind, in English, available publicly. However, our intended formulation also required the target image (reference image, pictorial summary). Therefore, we decided to extend the M3LS dataset by collecting the target images instead of creating a new dataset from scratch.

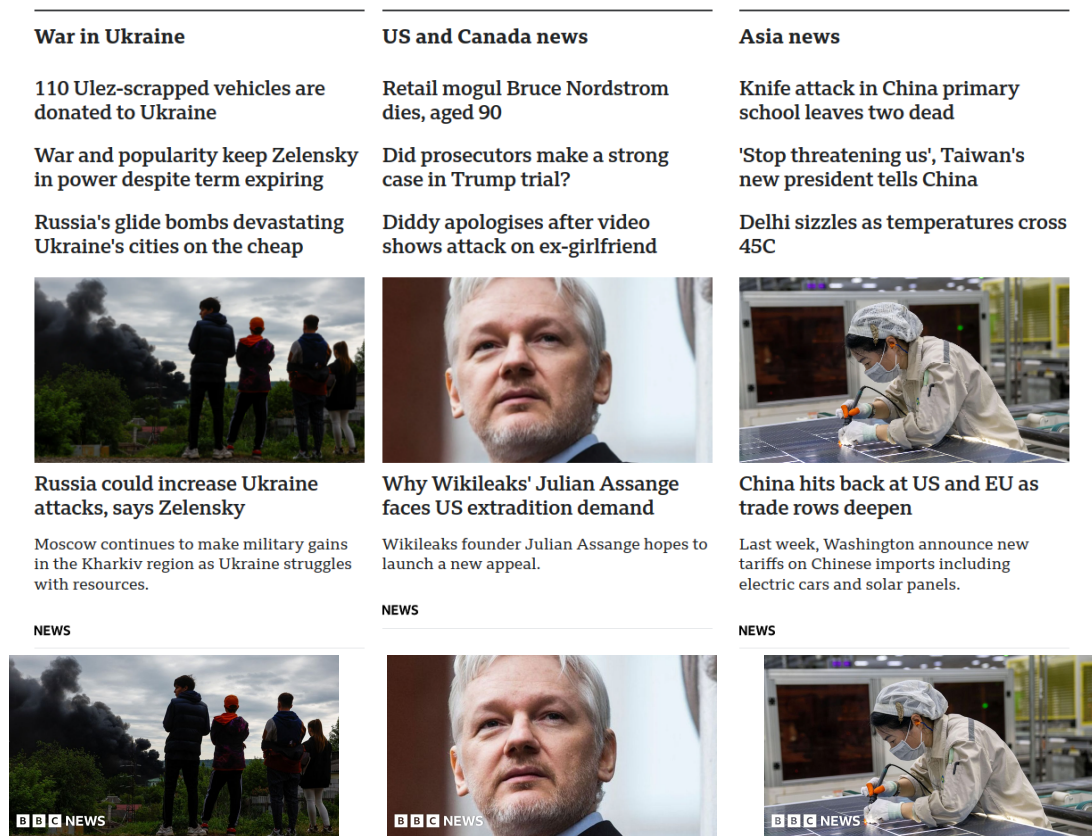


Figure 4.3: Above – a partial snippet of the [bbc.com](https://www.bbc.com) news hosting website. Below – the reference images (pictorial summaries) collected based on the specific HTML tag (see Section 4.3).

We started by analyzing the data instances of M3LS, as shared by the authors⁶. The dataset is shared in a *raw* form, with the textual articles encoded as `.json` files (see Figure 4.4). By analyzing the HTML structure of the website hosting the articles (`url` field, as provided in the `.json` files), we were able to

⁶<https://github.com/Raghvendra-14/M3LS>

identify a meta element characterized by a specific HTML tag (`property="og:image"`) that points to the reference image – the one representing the articles at the `www.bbc.com` main page (see Figure 4.3). After collecting the URLs pointing to the reference images (success rate of roughly 99.5%), we downloaded and deduplicated the pictures. The deduplication step was required, as a number of articles shared a common, template-like image. After this step, we were left with 210,071 articles (56%). Next, we applied a post-processing step that cropped the images to remove a watermark (see Figure 4.3). We considered this a necessary step, as the presence of a watermark would be an artificial indicator for a model that could skew the results. Finally, we filtered those multimodal articles from the M3LS dataset that fulfilled two conditions: they had at least a single image in the input and we were able to collect the target image for them, ending up with 115,432 instances. Following the image-based **MSMO** formulation, we appended the target image to the source images, with quantitative statistics of the number of input images in the extended M3LS dataset displayed in Table 4.4.

For the sake of future work and to establish a reproducible baseline, we proposed a split into training/validation/testing based on the publication date: articles published in January–April of 2021 for validation (5,865 instances), the ones published in May–October of 2021 for testing (6,854 instances) and the remaining ones (before January 2021) intended for training (102,713 instances).

We have released the final list of reference image URLs paired with IDs allowing to match with the original M3LS data instance via GitHub, i.e., <https://github.com/ufal/UNMHG>.

Min	Q_1	Mean	Q_3	Max
2	2	3.79	4	21

Table 4.4: Quantitative statistics of the number of input images (including the target image) in the subset of the English M3LS dataset that we extended with the multimodal target. Table reprint from Krubiński and Pecina (2024).

```

1  {
2    "url": "https://www.bbc.com/news/entertainment-arts...",
3    "title": "Time's Up: Boss quits over ties ...",
4    "date": "2021-08-27T09:48:22.000Z",
5    "summary": "Tina Tchen, the head of anti-sexual ...",
6    "0": {
7      "para": [
8        "Ms Tchen, a lawyer and the former chief ...",
9        "The high-profile politician resigned ...",
10       ...
11      ],
12      "images": [
13        [
14          "...f1d0e281821##0##1.jpg",
15          "Mr Cuomo continued to deny the allegations ..."
16        ],
17        ...
18      ]
19    },
20    "keyword": [
21      "#TimesUp campaign"
22    ],
23    "related": [
24      "https://www.bbc.com/news/world-us-canada-58153726",
25      ...
26    ]
27  }

```

Figure 4.4: An example of a data instance from the M3LS dataset. For simplicity, all of the retrieved fields are presented, but the actual content is truncated.

5. Experiments

In this Chapter, we will provide an overview of our experiments related to Multimodal Summarization, as published in Krubiński and Pecina (2023) and Krubiński and Pecina (2024). Our main goal is to provide the justifications and motivations for particular modeling choices and to introduce our qualitative and quantitative findings.

5.1 MLASK – MMS

This section is based on the MLASK: MULTIMODAL SUMMARIZATION OF VIDEO-BASED NEWS ARTICLES (Krubíński and Pecina, 2023) article.

5.1.1 Motivation and Overview

Our work on the video-based MSMO (VMSMO, as introduced in Section 2.2) develops from the insight that previous works explored mostly models trained from scratch on the limited, task-specific data. While the particular components, e.g., the image/frame feature extractor or the textual encoder, are initialized with weights from a generic, pre-trained checkpoint, the fact that there are numerous datasets and models available for the core task of text-only summarization was not explored deeply. Analyzing the experiments of Yu et al. (2021) (text-video input, but text-only output), we noticed that with the multimodal (visual) clues included in the input, the reported improvements were more significant for the models (variants) that performed worse in the text-only settings. Our intuition was that the benefits of the multimodal input as compared to the text-only one may be over-estimated by using too weak baselines. Furthermore, we observed the tendency to take the visual features for granted and focus on the modeling, i.e., not exploring alternative or combined feature extractors. In that regard, our goal was to conduct experiments with the same architecture but diverse visual features and see how it affects the overall quality. Eventually, we wanted to take a look at the robustness of a trained model – which elements can we simplify (ideally, at the runtime) without degrading the performance?

However, to conduct such experiments, we need access to the “raw” data and the “raw” model, and, as discussed in Section 4.2, at the beginning of our experiments, there was no video-based dataset and no VMSMO model¹ available publicly.

Therefore, our first step was to curate a video-based dataset (see Section 4.2). Having direct access to the data, we designed and implemented a multi-modal summarization model (MMS) that, thanks to its modular composition, allowed us to investigate our hypotheses and research questions.

¹The code-base of Li et al. (2020d), i.e., <https://github.com/irisxcy/VMSMO>, was public at that time, but it was not directly applicable to our needs. The trained model (checkpoint) was not shared by the authors.

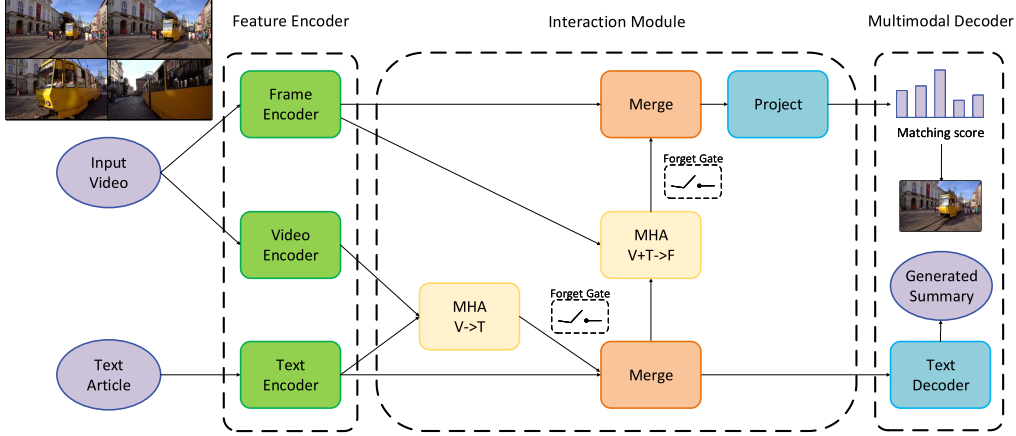


Figure 5.1: An overview of the proposed MMS model for Multimodal Summarization. Figure reprint from Krubiński and Pecina (2023).

In the remainder of this section, we go over the particular components of the base MMS model.

Overview

In our experiments, a video-based news article is represented by a pair (V, X) . V corresponds to the video input – a sequence of frames: $V = (v_1, v_2, \dots, v_N)$. X is the news article presented as a sequence of tokens: $X = (x_1, x_2, \dots, x_M)$. We assume that for each article, there is a ground-truth textual summary $Y = (y_1, y_2, \dots, y_L)$ and a ground-truth cover picture P . The task is to generate a textual summary \hat{Y} that includes the main points of the article and to choose a frame \hat{v} to act as a cover picture (pictorial summary). The proposed MMS model (see Figure 5.1) is structured into three parts: *Feature Encoder* composed of a text, video, and frame encoder, *Cross-modal Interaction Module* fusing the visual and textual representations, and *Multimodal Decoder* responsible for the summary generation and frame selection.

Feature Encoder

The Feature Encoder consists of a Text Encoder, a Video Encoder, and a Frame Encoder.

Text Encoder We use the Transformer encoder model to map the textual news article into the sequence of contextualized token embeddings (Equation 5.1).

$$X_{enc} = \text{TransformerEncoder}(X) \quad (5.1)$$

Video Encoder The news videos in our dataset are several minutes long and consist of hundreds of frames. To incorporate the short-term temporal dependencies, we segment the video into non-overlapping sequences of frames and use the 3D ConvNet for feature extraction (Equation 5.2). As the feature extractors, we use the model trained by Ghadiyaram et al. (2019) for video action recognition

on weakly-supervised social-media videos (internal data) and the visual component of the Text-Video model trained in a self-supervised manner by [Miech et al. \(2020\)](#) on the HowTo100M dataset ([Miech et al., 2019](#)). By default, we use a combination of both features by concatenating them along the hidden dimension. To incorporate the long-term temporal dependencies, we process the sequence of video features with the Transformer encoder model (Equation 5.3).

$$V_{enc} = 3D-CNN(V) \quad (5.2)$$

$$V_{enc} = \text{TransformerEncoder}(V_{enc}) \quad (5.3)$$

Frame Encoder To model the choice of a specific frame as a cover picture, frame-level representations are needed. In our experiments, we pre-process the input video by sampling one of every 25 frames as the cover picture candidates (1 frame per second). We examine the usage of EfficientNet ([Tan and Le, 2019](#)) and Vision Transformer ([Dosovitskiy et al., 2021](#)) as feature extractors and combine both by concatenating them along the hidden (feature) dimension. Both models were trained for image classification on the ImageNet ([Russakovsky et al., 2015](#)) dataset. To put the representations into context, we process the sequence of frame features with the Transformer encoder model (Equation 5.5).

$$V_{frame} = \text{CNN}(\text{Sample}(V)) \quad (5.4)$$

$$V_{frame} = \text{TransformerEncoder}(V_{frame}) \quad (5.5)$$

Before applying the Transformer encoder to visual features (Equation 5.3 and Equation 5.5), we project both the video and frame features into the same dimension as the hidden states of the text encoder.

Interaction Module

Following [Yu et al. \(2021\)](#), who examined different ways of injecting visual information into pre-trained generative language models, we employ the multi-head attention² (MHA) based fusion (see Section 2.1.2) to obtain the vision-guided text representation and perform the fusion after the last encoder layer (Equation 5.6–5.9).

$$Q = X_{enc}W_q; Q \in \mathbb{R}^{M \times d} \quad (5.6)$$

$$K = V_{enc}W_k; K \in \mathbb{R}^{N' \times d} \quad (5.7)$$

$$V = V_{enc}W_v; V \in \mathbb{R}^{N' \times d} \quad (5.8)$$

$$\tilde{X}_{enc} = \text{MHA}(Q, K, V); \tilde{X}_{enc} \in \mathbb{R}^{M \times d} \quad (5.9)$$

As suggested by [Liu et al. \(2020\)](#), we use the forget gate mechanism (Equation 5.10–5.11) so that the model can filter out low-level cross-modal adaptation information, and to provide a straightforward path to the encoded text representation X_{enc} . The \otimes symbol refers to a position-wise multiplication.

²We use the common notions/abbreviations for Transformer components, as introduced in [Vaswani et al. \(2017\)](#).

$$F = \text{sigmoid}(\text{Concat}(X_{enc}, \widetilde{X}_{enc})W_f) \quad (5.10)$$

$$\widehat{X}_{enc} = \text{Concat}(X_{enc}, F \otimes \widetilde{X}_{enc})W_{f'} \quad (5.11)$$

$$W_f, W_{f'} \in \mathbb{R}^{2M \times d}, \widehat{X}_{enc} \in \mathbb{R}^{M \times d}$$

We use the same MHA mechanism to obtain the text+video guided frame representations \widetilde{V}_{frame} by substituting X_{enc} with V_{frame} in Equation 5.6 and V_{enc} with \widehat{X}_{enc} in Equation 5.7 and Equation 5.8. The forget gate mechanism is applied to derive the final frame representations \widehat{V}_{frame} by substituting X_{enc} with V_{frame} in Equation 5.10 and \widetilde{X}_{enc} with \widetilde{V}_{frame} in Equation 5.11.

Multimodal Decoder

To generate the textual summary, we use the standard Transformer decoder with the vision-guided text representation \widehat{X}_{enc} as the input (Equation 5.12) and compute the standard Cross-Entropy loss (CELoss) w.r.t. the target sequence Y (Equation 5.13).

$$\widehat{Y} = \text{TransformerDecoder}(\widehat{X}_{enc}) \quad (5.12)$$

$$\mathcal{L}_{text} = \text{CELoss}(\widehat{Y}, Y) \quad (5.13)$$

To obtain the target labels C for the cover picture (cover frame) selection, we compute the cosine similarity between the numerical features of the target image and the candidate frames. The similarity of over 99.99% of instances was in the $[0,1]$ range, and the remaining negative values were mapped to 0. Following the previous works (Li et al., 2020d; Fu et al., 2020), we regard the frame with the maximum cosine similarity as ground-truth (C_{max}) and the others as negative samples, i.e., training with a binary signal. We use a projection matrix to map the text+video guided frame representations \widehat{V}_{frame} to a single vector (Equation 5.14) and compute the binary Cross-Entropy loss (BCELoss, Equation 5.15) w.r.t. target labels C . We train the whole model end-to-end by minimizing the sum of losses \mathcal{L} (Equation 5.16).

$$\widehat{C} = \widehat{V}_{frame}W_p; W_p \in \mathbb{R}^{d \times 1} \quad (5.14)$$

$$\mathcal{L}_{image} = \text{BCELoss}(\widehat{C}, C) \quad (5.15)$$

$$\mathcal{L} = \mathcal{L}_{text} + \mathcal{L}_{image} \quad (5.16)$$

5.1.2 Implementation

Models

We implement our experiments in PyTorch³ and use the small⁴ variant (300M trainable parameters) of the mT5 model, as provided via the Transformers (Wolf

³<https://github.com/pytorch/pytorch>

⁴<https://huggingface.co/google/mT5-small>

et al., 2020) package to initialize the textual encoder and the textual decoder. Following Yu et al. (2021), we use two separate 4-layer encoders to contextualize the video and the frame representations (Equation 5.5 and Equation 5.3).

As video feature extractors, we use the R(2+1)D 34-layer IG-65M⁵ and the S3D_HowTo100⁶ models to encode sequences of the length of 32 frames. i.e., to extract a single vector to encode every 32 frames. To extract frame-level features, we utilize the B5⁷ variant⁸ of EfficientNet and the vit-base-patch32-224-in21k⁹ variant of Vision Transformer to transform every frame into a single numerical vector. We follow the suggested pre-processing (e.g., image re-sizing, RGB channel normalization) for each feature extractor independently. The total number of trainable parameters is approximately equal to 323M. When computing the cosine similarity between the frame-level features, we compute the similarity with respect to both image-level feature extractors¹⁰ and average them to obtain the final similarity scores explored during training (see Section 5.1.1 and Section 5.1.3) and evaluation¹¹.

Data

In our experiments, we perform the training/dev/test splits of the MLASK dataset (see Section 4.2) following the chronological ordering based on publication date. We use the articles published in the first half (Jan–Jun) of 2021 for validation (2,482 instances) and the ones published in the second half (Jul–Dec) of 2021 and the beginning (Jan–Feb) of 2022 for testing (2,652 instances). The remaining data is used for training (36,109 instances).

For pre-training (see Section 5.1.3), we explore the large-scale, text-only, Czech news summarization corpus SumeCzech (Straka et al., 2018). SumeCzech was created by filtering the data from the Common Crawl project¹², based on a hand-picked list of popular Czech news websites, and consists of 1,001,593 articles. For each article in the dataset, the following were collected: the article’s headline (title), the article’s abstract, and the article’s text. Some additional metadata includes information such as the original URL or the article’s publication date. The authors proposed a data split based on the clustering of the embedded abstracts, dividing the data into training/dev/test/out-of-domain test as follows: 867,596/44,567/44,454/44,976. In our experiments, to avoid any training/test data leaks between MLASK and SumeCzech, we post-processed the training split of SumeCzech data by filtering out the articles that could¹³ appear in MLASK

⁵<https://github.com/moabitcoin/ig65m-pytorch>

⁶https://github.com/antoine77340/S3D_HowTo100M

⁷While working on the dissertation, we realized that in the conference paper, i.e., Krubiński and Pecina (2023), we mistakenly reported that the B4 variant was explored. In the public code-base, the correct model is referenced.

⁸<https://huggingface.co/google/efficientnet-b5>

⁹<https://huggingface.co/google/vit-base-patch32-224-in21k>

¹⁰Unless only one of the feature extractors is employed in the modeling process (see Section 5.1.3). In that case, there is a single similarity score.

¹¹Please consult the public repository for further technical details, i.e., <https://github.com/ufal/MLASK/>.

¹²<https://commoncrawl.org/>

¹³One of the news websites that were explored to curate MLASK, was a part of the hand-picked list based on which SumeCzech was created.

based on the date of publication (794,018 left, i.e., 92%).

Metrics

Most existing implementations of ROUGE are English-specific and utilize, e.g., an English stemmer or a word bank of English stop words. Since the MLASK dataset is in Czech, we follow the work of [Straka et al. \(2018\)](#) and evaluate the model performance with a language-agnostic $\text{ROUGE}_{\text{RAW}}$ ¹⁴ variant of ROUGE, reporting the F1 scores (ROUGE-1, ROUGE-2, and ROUGE-L). This variant utilizes no stemmer, no stop words, and no synonyms, tokenizing the (hypothesis/reference) texts based on white spaces.

To estimate the quality of cover frame selection, we report the cosine similarity (CosSim) between the reference image and the chosen cover frame. To examine the model performance besides the top-1 choice, we follow [Li et al. \(2020d\)](#) and report Recall@k¹⁵ (R@k) considering only the frame closest to the ground-truth as a positive example. To evaluate the frame scoring at even coarser, video-level granularity, we report Kendall’s τ ¹⁶ (KC) and Pearson’s r ¹⁷ (PC) correlation coefficients to measure the correlation of the ordering based on the projected representations (Equation 5.14) with the absolute frame ordering based on similarity with the ground-truth image.

Baselines

To put our experiments into a wider context, we report the performance of several text-only baselines: i) *RandomT* extracts three random sentences from the article; ii) *Lead3* extracts three initial sentences; iii) *Oracle* takes three sentences that maximize ROUGE-L with the ground-truth abstract (the upper bound for extractive summarization). Additionally, we report the performance of the (small) mT5 model fine-tuned for text summarization on the textual data in the Czech language – mT5-MLASK is fine-tuned on the textual part of the MLASK training set, and mT5-SumeCzech is fine-tuned on the SumeCzech (see Section 5.1.3) dataset. We also report a video-only baseline *RandomV*, which performs a random frame ordering.

Setup

We train the multimodal MMS model using the Adam optimizer ([Kingma and Ba, 2015](#)) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We increase the learning rate linearly for the first 8,000 steps (0 to 5e-4) and then follow an inverse square root decay schedule. Since both the text encoder and the decoder are pre-trained, we freeze them for the first 2 epochs. We limit the document size to 1,536 sub-word tokens and the summary length to 256 tokens. We train all the models for 50 epochs with an early stopping applied if ROUGE-L does not improve on the dev-set for 5 consecutive epochs. During decoding, we use the best checkpoint with respect to

¹⁴<https://lindat.cz/repository/xmlui/handle/11234/1-2615>

¹⁵<https://lightning.ai/docs/torchmetrics/stable/retrieval/recall.html>

¹⁶<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html>

¹⁷<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html>

DEV	ROUGE-1	ROUGE-2	ROUGE-L	CosSim	R@5	R@10	KC	PC
<i>RandomT</i>	13.92	1.63	9.02	-	-	-	-	-
<i>Lead3</i>	15.47	2.32	10.25	-	-	-	-	-
<i>Oracle</i>	22.92	5.37	18.28	-	-	-	-	-
mT5-MLASK	18.25	4.14	13.07	-	-	-	-	-
mT5-SumeCzech	19.18	4.53	13.76	-	-	-	-	-
<i>RandomV</i>	-	-	-	0.335	0.092	0.182	0.000	0.000
MMS	18.34	4.12	13.26	0.563	0.206	0.339	0.303	0.465
+ Masked Video	17.70	3.84	12.81	0.548	0.191	0.320	0.275	0.439
- IG-65M	17.74	3.89	12.95	0.558	0.200	0.323	0.290	0.456
- S3D	17.82	3.88	12.93	0.530	0.187	0.321	0.260	0.428
- Effnet	18.07	4.04	13.13	0.589	0.160	0.280	0.211	0.328
- ViT	17.69	3.71	12.82	0.527	0.192	0.320	0.309	0.488
+ SumeCzech	19.64	4.95	14.32	0.551	0.192	0.319	0.274	0.440
+ Smooth Labels	19.73	4.97	14.34	0.562	0.202	0.332	0.295	0.458
+ Masked Video	19.74	5.02	14.34	0.561	0.197	0.331	0.290	0.452
TEST								
MMS	18.45	4.29	13.42	0.552	0.183	0.321	0.306	0.447
+ Masked Video	17.65	3.95	12.88	0.542	0.187	0.332	0.283	0.422
- IG-65M	17.81	4.02	13.07	0.548	0.186	0.321	0.296	0.437
- S3D	17.89	4.03	13.03	0.531	0.177	0.316	0.264	0.408
- Effnet	18.21	4.28	13.37	0.582	0.157	0.279	0.216	0.311
- ViT	17.78	3.94	13.00	0.509	0.176	0.311	0.303	0.452
+ SumeCzech	19.58	4.95	14.30	0.541	0.181	0.318	0.278	0.420
+ Smooth Labels	19.74	4.90	14.34	0.551	0.188	0.330	0.299	0.444
+ Masked Video	19.69	4.91	14.38	0.553	0.184	0.326	0.300	0.439

Table 5.1: Evaluation on the dev-set and test-set of MLASK. The figures are averaged over three runs with different seeds. The three highest-scoring systems in each column are bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2023).

ROUGE-L, utilizing beam search with the beam size of 4, length penalty of 1.0, and repetition penalty (Keskar et al., 2019) of 2.5. We select the cover frame by applying argmax to the projected representations (Equation 5.14). We employ gradient accumulation to train with the effective batch size of 32. Each model is trained on a single GeForce RTX 3090 GPU, and the average training time is roughly 36 hours. The text-only models (see Section 5.1.3) are fine-tuned with the Adafactor (Shazeer and Stern, 2018) optimizer, using a constant learning rate equal to 5e-4 and trained until ROUGE-L ceased to improve on the relevant dev-set for 5 consecutive evaluations.

5.1.3 Results and Ablation Studies

A sample of model outputs is provided in Appendix B.2.1.

Looking at the results of the baseline systems (see Table 5.1), both text-only *mT5* variants outperform the trivial baselines (*RandomT*, *Lead3*), but their results are below the *Oracle* performance. Using larger training data (SumeCzech has roughly 20 times more documents than MLASK) improves the performance

by approximately 1 ROUGE point. As expected, with the *RandomV* baseline, the value of R@10 is twice the value of R@5 – with a random frame ordering, doubling the set of top-k positions increases the probability of including the single positive frame (see Section 5.1.1) twice.

The baseline multimodal system (MMS) achieves slightly higher scores than mT5-MLASK (dev-set ROUGE-1: 18.25 \rightarrow 18.34, ROUGE-L: 13.07 \rightarrow 13.26) but lags behind the text-only mT5-SumeCzech that was trained on a much larger corpus. While the exact CosSim values can not be compared (they are based on different feature extractors), the Δ CosSim between the trained model and the random baseline is much more significant compared to the work of Fu et al. (2021) and our unified, multi-task formulation (see Section 5.2.3).

In order to understand the effect of the particular components and design choices, we propose a number of ablation studies related to image/frame representations, task-specific pre-training, and the formulation of the training signal.

Firstly, to analyze the effect of the individual visual features (see the discussion on Video Encoder and Frame Encoder in Section 5.1.1), we report the results of the MMS model, excluding those features one by one (see the rows starting with the “–” sign in Table 5.1). The scores indicate that the model combining all the features (i.e., MMS) is superior, as demonstrated by the higher ROUGE score and the higher values of R@5 and R@10. The variant without the Effnet frame (image-level) features (MMS_{-Effnet}) achieves higher CosSim, but the value of CosSim is computed based on different numerical representations, and thus, the comparison is not straightforward. To avoid this issue, we recommend that future works disentangle image/frame representations used for computation from image/frame representations used to compute the similarity. Compared to the remaining variants with certain features excluded (that all perform comparatively), the MMS_{-Effnet} variant achieves higher ROUGE values but lower values of correlation coefficients. This can be explained by the difference in the size of the extracted feature vectors, i.e., 512/512/2,048/768 for IG-65M/S3D/Effnet/ViT.

Secondly, we wish to establish the role of pre-training. When discussing previous works on Multimodal Summarization in Chapter 2, we remarked that a number of works that report significant improvements (Δ ROUGE of 3-4) in terms of the quality of textual output when comparing text-only with multimodal models, train their models from scratch. Within the text-centric formulation, since the textual document is available in the input, it is, in principle, possible to generate a high-quality summary based only on the textual input¹⁸, ignoring the visual clues. Therefore, from the modeling perspective, if we train the model end-to-end with the training signal based on the textual output, we can not guarantee that the model will pay a lot of attention to the visual clues. Our hypothesis is as follows: the better the model is at text-only summarization, the less effective the usage of visual clues will be, as the model will focus on the textual content. On the other hand, since the textual summary is part of our desired output (also in the MSMO formulation), we wish to train our models following the established

¹⁸This “greedy” characteristic of multi-modal deep neural networks is nicely discussed in Wu et al. (2022), and the follow-up works.

pipelines, including the task-specific pre-training.

To transform this discussion into numbers, we pre-train the textual component of the MMS model (mT5) on text-only summarization using the SumeCzech corpus. In the next step, we replicate the multimodal training and report the performance in Table 5.1, i.e., $\text{MMS}_{+\text{SumeCzech}}$. The results are consistent across dev-set and test-set – we observe an improvement in the quality of textual output (ΔROUGE of 1.0-1.5), with a slight degradation in terms of the visual quality, consistent along all metrics. To deepen this study, we considered using the mT5 architecture for the core textual component but initializing the weights with random values. Theoretically, this could provide further insight into the capability of consuming visual clues. However, we finally decided against performing such an experiment. When working with fine-tuned models, one can build upon the experience of others and, to a certain degree, ensure that the correct training hyper-parameters are applied and focus on their problem/task. When training from scratch, we would need to perform a large-scale hyper-parameter tuning. Otherwise, we would not be able to confidently say that the difference in performance is due to the usage of visual clues, as opposed to applying an ineffective training regime.

Thirdly, we look at the training signal from a visual perspective. In Section 5.1.1, we explained that to obtain the target labels C for the cover picture selection, we compute the cosine similarity between the numerical features of the target image and the candidate frames. We transfer those similarities into a binary signal C_{max} by regarding the frame with the maximum cosine similarity as a positive and the remaining ones as negatives.

After examining the cosine similarity patterns (see Figure 5.2 and Figure 2.4), we made an observation regarding the per-video similarities. We noticed that one of two things may happen – either there is more than one peak, or there are consecutive sequences of frames with very similar scores, i.e., capturing a still scene. Our intuition was that this might harm the model performance by introducing noise during training – very similar frames might be labeled as both positive and negative examples. To overcome this issue, besides the binary labels C_{max} , we introduce the smooth labels C_{smooth} that assign to each frame its “raw” cosine similarity score with the target image and re-train the MMS model with those smooth target labels (the “+ Smooth Labels” variant¹⁹). The difference in terms of metric values is minimal but consistent across the dev-set and the test-set and across (almost) all of the metrics that we consider.

Finally, we perform an experiment to probe the models for sensitivity with respect to the visual features. Namely, we mask the video features (in the Video Encoder) with random noise during both training and evaluation. The model can still access the visual information via the image-level frame features (in the Frame Encoder), which are left intact.

Surprisingly, for the variant that was pre-trained on a large text-only corpus ($\text{MMS}_{+\text{SumeCzech}}$), masking the video features does not hurt the model performance, resulting in even slightly higher scores, i.e., ROUGE-L: 14.34 \rightarrow 14.38,

¹⁹To linearize our research, we decided to perform the experiments with smooth labels based on the model pre-trained for text-only summarization.

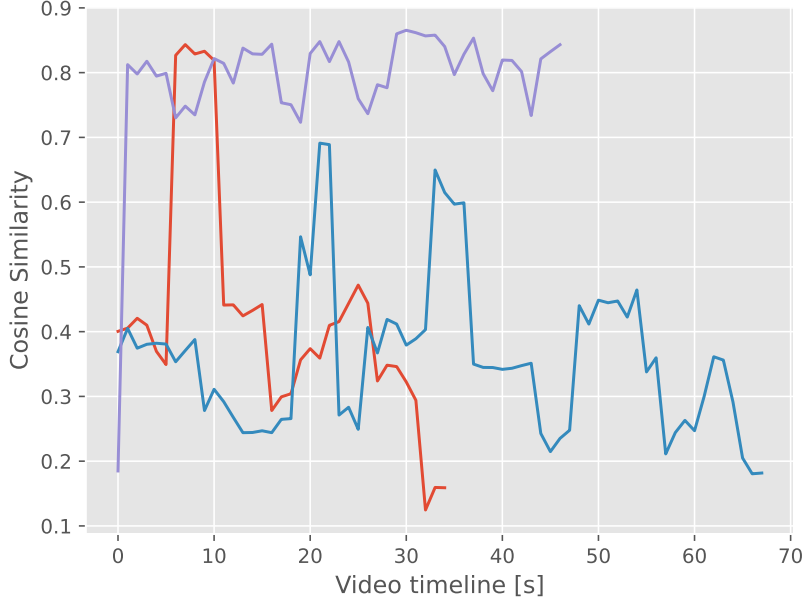


Figure 5.2: Three examples of cosine similarity (y-axis) plots between the numerical features of the reference cover picture and all candidate frames (x-axis) from the video. The examples were chosen manually from the MLASK dataset to present three different video similarity patterns: with a single peak (red), with more than one peak (blue), and with a consecutive sequence of frames having very similar scores (violet). Figure reprint from Krubiński and Pecina (2023).

	Total Score	Adequacy Score
<i>Reference</i>	2.89 ± 0.99	1.64 ± 0.50
<i>Random V</i>	2.39 ± 1.15	1.44 ± 0.61
System A	2.64 ± 1.10	1.51 ± 0.58
System B	2.66 ± 1.04	1.56 ± 0.52

Table 5.2: System performance on the task of cover picture selection, validated on the subset of the MLASK test-set. Table reprint from Krubiński and Pecina (2023).

CosSim $0.551 \rightarrow 0.553$, as reported on the test-set. However, the quality drop is observed for the model that did not go through the task-specific pre-training, i.e., ROUGE-L: $13.42 \rightarrow 12.88$, CosSim $0.552 \rightarrow 0.542$, as reported on the test-set. After examining the models, we noticed that the representations after the video encoder (Equation 5.5) are not very meaningful, i.e., every segment is mapped to a similar vector. We believe this is due to the indirect usage of video representations in the Cross-modal Interaction Module – too weak of a learning signal (gradient) is propagated to the video encoder. Those observations are in line with our findings related to pre-training – the “weaker” is the core textual component, the more effective are the additional (visual) clues in the input.

Since those results were rather counter-intuitive, we decided to perform a human evaluation to validate the quality of the pictorial output. Due to the lack of a standardized evaluation framework, we designed one ourselves (see Section 3.3), focusing only on the pictorial output. We decided to validate the outputs from

two systems – MMS pre-trained on SumeCzech using the smooth labels (MMS + SumeCzech + Smooth Labels, further denoted as System A) and the same model with masked video features (MMS + SumeCzech + Smooth Labels + Masked Video, further denoted as System B). For comparison, in the evaluation, we also included the reference picture and a random frame from the video. The annotators were shown the outputs from several systems at once, with the textual context provided via the article’s title and the reference summary. They were asked to rate the images on a Likert scale (see Section 3.1.1) of 0 to 4 (the higher, the better). The levels on the Likert scale were designed to consider both the relevance and the quality – it may happen that a particular frame sampled from the video is blurred, as it may resemble e.g., a fading away end of a scene.

The system-level averages of the scores assessed by the human annotators (Total Score) are reported in Table 5.2. On average, the reference picture is assigned the highest score, and our proposed multimodal summarization model performs better than the random baseline. The results of human assessment confirm our previous findings based on automatic metrics – that the model is not utilizing the video features in an effective manner. It is worth noticing, however, that even the reference picture is not considered very relevant (average score below 3) and that none of the differences are statistically significant. To examine the stability of the annotation process, we also report the averages (Adequacy Score) that disregard the quality of the image and focus only on the relevance. We do this by mapping the labels from Section 3.1.1, (i.e., $0 \rightarrow 0$; 1 and $2 \rightarrow 1$; 3 and $4 \rightarrow 2$). The results are in line with the original ones.

5.1.4 Implications

Since the publication of the MLASK dataset and the code-base related to our experiments, a number of follow-up works have been published. Some of them (e.g., Shohan et al. (2024); Bao et al. (2024)) mention our contribution by using it as an example of a video-based problem and put it into a wider context by comparing it with other recent advances. However, (at least) two works directly apply the MMS architecture to their respective problems/tasks.

Qiu et al. (2024) use our formulation of Feature Encoder, with separate Video Encoder and Frame Encoder, exploring the same visual feature extractors. Similarly, for thumbnail generation (frame selection) they use the same formulation of Multimodal Decoder. The authors extend our work by substituting the Text Encoder with a hierarchical one. In their experiments²⁰, the length of the text input is, on average, longer. Therefore, they firstly perform a sentence-splitting step and extract a single vector (corresponding to the special [CLS] token) from each sentence. Then, the sentence representations are contextualized together with a second Transformer-based text-only encoder. Additionally, the authors improve over our uniform sampling of input frames. Namely, they explore the differences of adjacent frames (represented as CNN feature vectors) to define scenes in the video, setting a particular threshold to draw scene boundaries. Using K-means and Euclidean distance, they cluster the candidate frames per scene and remove redundant (semi-duplicate) candidates from each consecutive scene.

²⁰https://github.com/Jason-Qiu/MMSum_model

Faheem et al. (2024), who introduce the video-based UrduMASD dataset in Urdu, claim to use the off-the-shelf version of the MMS model as the sole multimodal system in their experiments. Since their code-base is not public, we are not able to say whether any modifications were applied.

5.2 Unifying Uni- and Multi-modal Summarization

This section is based on the TOWARDS UNIFIED UNI- AND MULTI-MODAL NEWS HEADLINE GENERATION (Krubiński and Pecina, 2024) article.

5.2.1 Motivation and Overview

This work builds upon the observation that the recent approaches to modeling the task of Multimodal Summarization explore sophisticated, modular architectures built upon hierarchical cross-modal encoders and modality-specific decoders, which restrict the model’s applicability to specific data modalities – once trained on, e.g., *text+video* pairs there is no straightforward way to apply the model to *text+image* or *text-only* data. From the modeling perspective, there are a number of issues that one must solve:

- In order to accept both text-only and multimodal input, there must be a unified approach to generating the encoded input representation so that they end up in the same subspace. Otherwise, the textual decoder would need to generate texts conditioned on representations from separate partitions/clusters. The visual inputs are supposed to act as additional, helpful clues, but the final, encoded multimodal representation should be close to the text-only representation in the space of all encodings.
- Designing an architecture that can effectively accept a varying number of images/frames in the input. If each input image i is encoded as a sequence²¹ $tt = tt(i) \in \mathbb{R}^{197 \times 768}$, then, assuming 20 images in the input, a simple concatenation-based approach (see Section 1.2.2) would require the model to operate on at least $20 \times 197 = 3940$ input tokens. This is a few times more than the maximal input length used for training recent multimodal encoders (see, e.g., Xu et al. (2023c); Peng et al. (2023); Alayrac et al. (2022)).
- Handling the temporal dependency in videos. As discussed in Section 2.2, approaches based on 3D ConvNets were effectively used to encode the sequence of frames representing the video. However, if the input is a collection of images, we would like to restrain from any temporal dependencies – in the image-based variant, we assume that the input is a *set* of images.

The task gets even more complicated if we approach the MSMO formulation. Typical modeling techniques for MSMO (see Section 2.2) compute a numerical score $s_i \in \mathbb{R}$ for each input image/frame (based on the encoded representations) and via the use of softmax transform them into the probability distribution over all input images/frames, i.e., $\text{softmax}(s_{i \in I})$. Such a formulation allows us to either rank the input images/frames or simply pick the most likely one, depending on the applications – but it requires a dedicated scoring module. While the SOTA multimodal LLMs (OpenAI, 2024; Yin et al., 2024) are capable of solving a vast number of complex, ViL tasks, such as complicated Optical Character

²¹That would be the case when using the popular ViT-L/14 variant of CLIP, i.e., <https://huggingface.co/openai/clip-vit-large-patch14> as a feature extractor.

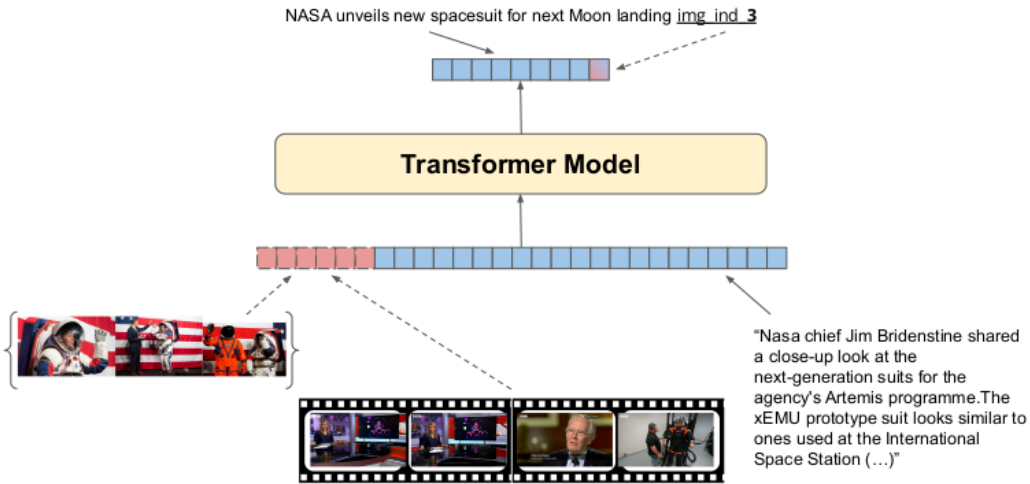


Figure 5.3: Overview of the proposed unified approach to **MSMO**. The visual tokens are appended to the text representation. The generated output includes the textual summary and the *index token* that indicates which input image (first, second, third, etc.) is picked as the pictorial summary. During training, a mixture of video-based, image-based, and text-only data is used. Figure reprint from [Kru-biński and Pecina \(2024\)](#).

Recognition (OCR) or geometry problems, we believe they still lack some capabilities required for solving the unified variant of Multimodal Summarization, as discussed above. Therefore, we propose a novel **UNMHG**²² formulation that utilizes a simple encoder-decoder model to summarize both uni- and multi-modal documents and introduce the *index tokens* to indicate which input image/frame (first, second, third, etc.) should be considered as the pictorial summary, allowing us to handle the multimodal output. For brevity, we follow the TL;DW formulation proposed by [Tang et al. \(2024\)](#) and use the article title as the textual target (i.e., the headline), although the proposed methods can also be applied for other summarization tasks, such as abstract generation.

An overview of the proposed formulation is presented in Figure 5.3. We transform the visual inputs into a sequence of image features and concatenate them with the textual (input) token embeddings. Visual features are contextualized together with textual process, i.e., there is no modality-specific encoder. Instead of using a dedicated module for image scoring, we realize the target image/frame representation by appending an *index token* to the textual target – `img_ind_1` indicates that the *first* image is the target, `img_ind_2` that the *second*, etc. This formulation allows us to use the standard Transformer architecture trained end-to-end in a multi-task setting – for the text-only input, we do not extend the textual embeddings and do not add the index token into the target sequence. Within the image-based formulation, we shuffle the input images during training to avoid positional bias.

The *index tokens* are inspired by the previous work on one-for-all architectures, unifying several vision-and-language tasks, such as the work of [Cho et al. \(2021\)](#) who introduced the visual sentinel tokens corresponding to image regions,

²²UNi- and Multi-modal News **H**eadline **G**enerattion

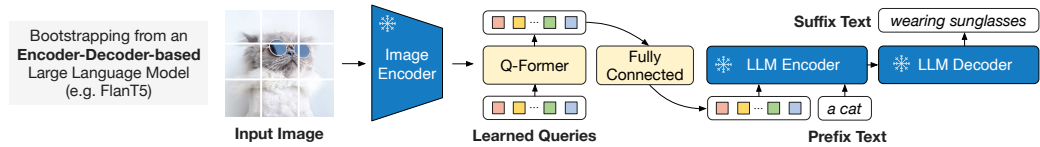


Figure 5.4: Overview of the BLIP-2 model, with an encoder-decoder LLM as the textual component. Figure reprint from Li et al. (2023).

allowing them to realize Visual Grounding with a text-only decoder, or the Task and Modality-Agnostic OFA framework (Wang et al., 2022c) that unified the multi-modal (only single-image inputs were considered) and text-only tasks with a sequence-to-sequence Transformer. In our experiments, we explore three input/output formats, i.e., $text+video \rightarrow text+image$, $text+images \rightarrow text+image$, and $text \rightarrow text$. Following Wang et al. (2022a), we handle the temporal dependencies in a simplified manner – sinusoidal positional embeddings are added to the visual tokens only when the visual modality comes from a video. We explore two modeling approaches: i) extending a text-to-text baseline with visual features and ii) fine-tuning a multimodal foundation model. Building upon the experience (Johnson et al., 2017) of multi-lingual MT, we append a string (" $t+v \rightarrow t+i$ ", " $t+i \rightarrow t+i$ ", " $t \rightarrow t$ ") to the input document, to act as a task indicator.

5.2.2 Implementation

Models

As the text-to-text baseline, we use the T5 (Raffel et al., 2020) v1.1 base²³ variant (250M trainable parameters) that we enrich with visual features extracted with frozen ViT-L/14 CLIP²⁴ (Radford et al., 2021) – we refer to this model as T5CLIP. We extract a single vector per image/frame²⁵ that we project with a linear layer to match the hidden dimension size of textual token embeddings. We extend the model vocabulary with index tokens, i.e., `«img_ind_1, img_ind_2, ...»` that are used for image/frame selection. We train the whole model end-to-end with the Adafactor (Shazeer and Stern, 2018) optimizer using the default parameters from the Transformers package²⁶.

For the multimodal baseline, we use the Flan T5-XL (Chung et al., 2023) version of BLIP-2²⁷ (Li et al., 2023, 3.9B parameters). In the BLIP-2 paper, the authors propose an efficient pre-training strategy to bootstrap a ViL model from an off-the-shelf frozen image encoder and frozen language model, with both encoder-decoder and decoder-only variants considered. Such an approach is motivated by data availability. Since there is much more image-only and text-only

²³https://huggingface.co/google/t5-v1_1-base

²⁴<https://huggingface.co/openai/clip-vit-large-patch14>

²⁵We use the “pooled” representation that extracts encoded embedding of the special CLS token. An alternative approach that averages the representations along patches (sequence dimension) is also possible, but we have not explored it. For details, please consult the Vision Transformer paper, i.e., Dosovitskiy et al. (2021) or see Section 1.2.1.

²⁶Please consult the public repository for further technical details, i.e., <https://github.com/ufal/UNMHG/>.

²⁷<https://huggingface.co/Salesforce/blip2-flan-t5-xl>

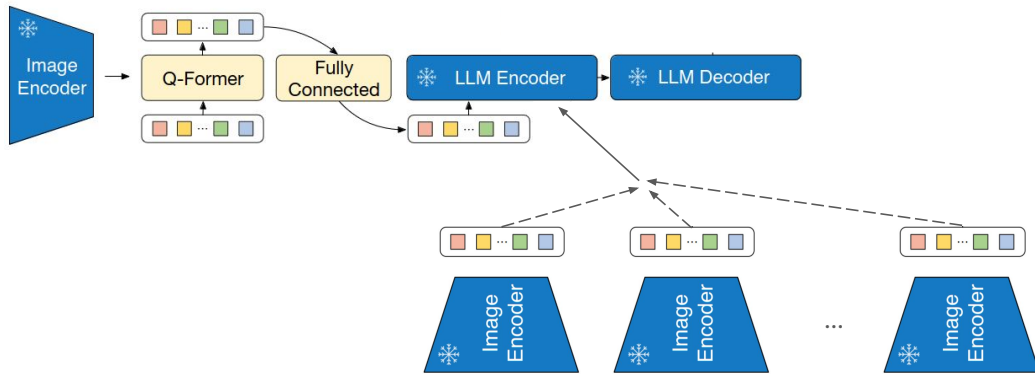


Figure 5.5: Overview of the BLIP-2 model extended to handle multiple images/frames in the input.

data compared to multimodal resources, the idea is to leverage the models pre-trained on a single modality and use the cross-modal data only to align the representations. The modality alignment is done via a novel Q-Former module, see Figure 5.4. Q-Former is a shallow Transformer network that uses the K and V sequences from the frozen image encoder and learns a fixed-sized query embeddings Q. The transformed image representation is projected with a linear layer to match the dimension of textual token embeddings and is finally concatenated with the textual token, before passing through the frozen LLM. With this approach, the image representation gets compressed (in the sequence dimension, the learnable query is much shorter than the original image representation, i.e., 32 vs 257 for the BLIP-2 variant that we explore), and the frozen LLM can digest the additional, visual information, similarly to the prefix-tuning (Li and Liang, 2021) approach. Once fine-tuned, the model weights can still be used for text-only inference by directly querying the frozen LLM component.

However, there is no straightforward way of applying the BLIP-2 model to an input with multiple images. Therefore, we propose a modification to allow such an application. Namely, we concatenate the Q-Former features from multiple images before appending them to the textual embeddings, introducing no new trainable parameters (see Figure 5.5). Fine-tuning the whole BLIP-2 is feasible neither from the computational perspective (memory requirements of GPU) nor from the design perspective (we wish to keep the main components frozen). However, since we alter the visual prefix (multiple images) and our intended task – Multimodal Summarization – is more specific than the generic pre-training of BLIP-2, a fine-tuning step is required. We decided to use the LoRA (Hu et al., 2022) procedure and update only the Q and V matrices in the Q-Former and Language Model components (5.7M trainable parameters in total), training with the AdamW (Loshchilov and Hutter, 2019) optimizer with $\beta=(0.9, 0.999)$, learning rate of $1e-5$ and weight decay of $5e-2$. Since the LoRA procedure prohibits us from updating the matrix of token embeddings, we do not add the index tokens directly into the model vocabulary but rely on the model to generate them from available components²⁸. We also simplify the initial design by removing the

²⁸Once fine-tuned, we use a simple regex (`img_ind_\d+`) to identify and extract the *index token* from the output. While evaluating on the test-set, the model properly generated a sequence corresponding to *index token* in 100% of the cases.

temporal dependency and treating sequences of frames and collections of images alike.

Data

In our experiments, we use the text-only PENS (Ao et al., 2021), the video-based MLASK, and the image-based M3LS datasets for training and testing.

The PENS dataset²⁹ contains 113,762 news articles in English and was originally introduced for personalized news headline generation. We filtered it by removing articles identified as non-English by the langid³⁰ language identifier, and those where the title has less than 2 words or more than 25 words. In the next step, we de-duplicated the data based on the article and title fields. We were left with 100,992 documents (89%), out of which 5,000 were used for validation and testing and the remaining ones (90,992) for training. Since the textual part of MLASK – at the time of our experiments, the largest publicly available video-based news summarization dataset – is in the Czech language, we used the CUBBITT (Popel et al., 2020) Machine Translation system³¹ to translate articles and summaries (titles) into English. We use the same data split as previously, i.e., 36,109/2,482/2,652 instances for training/validation/testing. As described in Section 4.3, we extend the English subset of the image-based M3LS dataset by collecting the cover pictures and use the data split based on the publication date that we proposed, i.e., 102,713/5,865/6,854 instances for training/validation/testing.

Metrics

We measure the quality of the textual output with ROUGE-L and BERTScore, reporting the F1 scores. For the pictorial output, we report the cosine similarity (CosSim) between the ViT-L/14 CLIP features of the target image and the one chosen by the model. To measure the multi-modal interactions, we report the CLIPBERTScore metric, adapted to the MSMO variant (see Section 3.3). It is computed as a weighted average³² of the CLIPScore of the chosen image and the generated summary and the BERTScore precision of the input article and the generated summary. For the image-based data, we also report the top-1 accuracy (Top-1 Acc), i.e., the percentage of predictions where the target image is correctly retrieved. We use the ROUGE metric from the TorchMetrics package³³ and the original implementations of BERTScore³⁴ and CLIPBERTScore³⁵. The signature of the BERTScore model that we use is: `roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.29.0.dev0)-rescaled`. For readability reasons, we re-scale both BERTScore and CLIPBERTScore into the [0–100] range by multiplying the numerical scores by 100.

²⁹https://msnews.github.io/pens_data.html

³⁰<https://github.com/saffsd/langid.py>

³¹<https://ufal.mff.cuni.cz/cubbitt>

³²We use the recommended $\alpha = 0.25$, i.e., $\text{CLIPBERTScore} = \alpha \times \text{CLIPScore} + (1 - \alpha) \times \text{BERTScore}$

³³https://torchmetrics.readthedocs.io/en/stable/text/rouge_score.html

³⁴https://github.com/Tiiiger/bert_score

³⁵<https://github.com/meetdavidwan/faithful-multimodal-summ>

Baselines

We report two textual, extractive baselines: *Lead* that extracts the first sentence and *Oracle* that picks a sentence maximizing ROUGE-L with the ground-truth. For the off-the-shelf textual abstractive baselines, we use the Alpaca (Taori et al., 2023) and BRIO (Liu et al., 2022b) models. The Stanford Alpaca model³⁶ is a text-only, Transformer-based LLM, fine-tuned from the LLaMA (Touvron et al., 2023a) model to follow instructions. It has been trained on the automatically generated data created with the Self-Instruct (Wang et al., 2023b) techniques. In our experiments, we use the following prompt:

```
Below is an instruction that describes a task, paired with an
input that provides further context. Write a response that
appropriately completes the request.
```

```
### Instruction:
```

```
Generate a one-sentence summary of a given text, using no more
than 10 words.
```

```
### Input:
```

```
__DOCUMENT_TEXT__
```

```
### Response:"
```

We report results with the 7B parameter variant and, for generation, utilize beam search of size 4, length penalty of -5.0, and repetition penalty of 2.5. In our early experiments, we noticed that truncating the input at the token level resulted in words and sentences being cut in half, which negatively affected the model performance. To avoid this, we use the `wtpsplit` package (Minixhofer et al., 2023) to prompt the model with full sentences, capping the input length (i.e., `__DOCUMENT_TEXT__`) at 1000 characters. BRIO (Liu et al., 2022b) is a recent encoder-decoder model trained for both summary *generation* and *evaluation*, i.e., the ability to score the quality of candidate summaries. We use the Yale-LILY/`brio-xsum-cased` variant (568M parameters), which is based upon the pre-trained PEGASUS (Zhang et al., 2020a) model and fine-tuned on the XSum (Narayan et al., 2018) dataset to generate single-sentence summaries.

For the video-based data, we compare with a variant of the MLASK-MMS (MMS) model (see Section 5.1). We use the configuration based on the small variant of mT5, with all four feature extractors and activated smooth labels, and train the model on the MLASK dataset machine-translated into English. We report a trivial baseline *Random Vi* that picks a random image/frame for both the video-based and image-based data. To further establish a comparison with the recent developments (see Section 2.2), we also report a generative visual baseline based on Stable Diffusion. We employ the `stabilityai/stable-diffusion-2-1` model prompted with the textual target (`_TEXT_`) using the following template: “High quality, photorealistic photo of `_TEXT_`”.

³⁶https://github.com/tatsu-lab/stanford_alpaca

Setup

We train all the models for up to 10 epochs with early stopping applied if ROUGE-L F1 does not improve for 5 consecutive epochs. We limit the source size to 1024 sub-word tokens and the target length to 128 tokens. We train on a machine with three NVIDIA A40 GPUs, and the average training time is 24 hours for the T5 variants (effective batch size 300) and one week for the BLIP-2 variant (effective batch size 60). During decoding, we utilize beam search of size 4, length penalty of 1.0, and repetition penalty of 2.5. Unless training on data corresponding only to a particular input/output format (e.g., $text+video \rightarrow text+image$), all of the trainings are done in a multi-task fashion. At each training step, a mini-batch is sampled from every dataset (PENS/MLASK/M3LS). The three mini-batches are processed one-by-one, and the average value of the loss is back-propagated to update the trainable parameters.

5.2.3 Results

A sample of model outputs is provided in Appendix B.2.2.

	ROUGE-L						BERTScore					
	MLASK		PENS		M3LS		MLASK		PENS		M3LS	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
<i>Lead</i>	12.28	12.19	16.51	16.27	9.74	9.85	10.67	10.77	8.85	9.10	9.57	10.03
<i>Oracle</i>	24.44	25.01	38.99	39.17	23.85	23.65	21.09	21.99	31.78	31.91	18.43	19.34
Alpaca	14.81	15.07	26.80	26.92	16.54	16.96	18.67	19.14	28.40	28.62	19.34	20.78
BRIO	15.56	15.58	16.40	16.55	18.18	18.79	15.97	16.49	16.61	16.83	23.30	25.03
T5CLIP _{MLASK}	20.79	21.32	-	-	-	-	25.46	25.99	-	-	-	-
T5CLIP _{PENS}	-	-	43.00	44.21	-	-	-	-	45.12	46.70	-	-
T5CLIP _{M3LS}	-	-	-	-	29.63	29.68	-	-	-	-	33.84	34.48
T5CLIP	21.48	21.43	43.07	44.47	29.64	29.38	26.43	26.36	45.24	46.80	33.16	33.73
T5CLIP _{w=10}	21.48	21.57	42.60	43.74	29.32	29.28	25.98	26.43	44.31	45.74	32.67	33.25
T5CLIP _{w=50}	20.63	21.05	40.87	42.15	26.92	26.88	25.21	25.55	41.72	43.40	29.14	29.71
T5CLIP _{Smooth}	21.30	21.32	43.25	44.39	30.06	30.03	26.50	26.24	45.53	46.94	33.70	34.44
BLIP-2	23.25	24.24	43.03	44.37	32.82	33.02	27.87	28.94	44.56	46.27	35.91	37.24
MMS	19.99	20.07	-	-	-	-	23.97	24.38	-	-	-	-

Table 5.3: Evaluation of the textual output quality on the validation and test splits for each modality-specific dataset (PENS for $text \rightarrow text$, MLASK for $text+video \rightarrow text+image$, and M3LS for $text+images \rightarrow text+image$). The three highest-scoring systems in each column are bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2024).

Textual Output

Table 5.3 compares the models trained separately on each task/dataset (e.g., T5CLIP_{PENS}) with the ones trained in the multi-task fashion (T5CLIP). The results are comparable, with additional textual data improving the performance on the smallest video-based dataset – MLASK. MLASK has roughly one-third of the documents compared to both PENS and M3LS, and since it was machine-translated into English, the addition of genuine textual data could have been expected to help. The proposed baselines are lagging behind the task-specific models. By manually examining the outputs of Alpaca, we notice that the model tends

to generate more than one sentence in the output. Despite explicitly instructing the model to generate the summary `...using no more than 10 words`, on average, there are 15.8 words in the summary, with the results consistent among all three datasets. To put this number into perspective, the headlines (titles) in the test-split of PENS/MLASK/M3LS have, on average, 10.5/9.1/12.3 words. Despite being trained with the single-sentence summaries from XSum as a target, BRIO generates, on average, as many as 29 words in the summary. Those results suggest that the headline/title/single-sentence summary may correspond to distinct concepts and might not be comparable between datasets – even from the same domain of news summarization.

Overall, the highest scores are obtained by the fine-tuned BLIP-2, which integrates the largest language component – Flan T5-XL. The findings based on both metrics considered (ROUGE-L and BERTScore) are aligned. The only major discrepancy is the *Oracle* baseline, which, however, by design, is biased towards ROUGE. The MMS model achieves the lowest scores, but the comparison is not completely fair – we have not pre-trained the core textual component for summarization, which in our experiments on MLASK was shown to be a crucial factor (see Section 5.1). If we compare the performance of the MMS model with the T5CLIP_{MLASK} one that was fine-tuned on the same data, the difference is not very significant, with Δ ROUGE-L of roughly 1. The difference can be attributed to either the size of the textual component – MMS uses a “small” variant of mT5, while T5CLIP_{MLASK} employs the “base” variant of T5, or the fact that the T5 component is English-only, and thus the vocabulary is better suited for English input/output.

	CosSim				CLIPBERTScore				Top-1 Acc	
	MLASK		M3LS		MLASK		M3LS		M3LS	
	dev	test	dev	test	dev	test	dev	test	dev	test
<i>RandomVi</i>	0.61	0.61	0.75	0.76	-	-	-	-	33.20	33.59
T5CLIP _{MLASK}	0.64	0.64	-	-	70.56	70.59	-	-	-	-
T5CLIP _{M3LS}	-	-	0.97	0.97	-	-	69.57	69.70	93.59	94.56
T5CLIP	0.64	0.64	0.93	0.94	70.67	70.65	69.61	69.77	87.49	88.55
T5CLIP _{w=10}	0.64	0.64	0.96	0.97	70.99	70.99	69.74	69.92	93.03	94.05
T5CLIP _{w=50}	0.64	0.63	0.96	0.97	71.12	71.11	69.60	69.72	91.76	93.19
T5CLIP _{Smooth}	0.64	0.63	0.82	0.81	70.65	70.61	69.83	69.96	39.91	38.55
BLIP-2	0.63	0.62	0.83	0.84	71.46	71.44	70.07	70.26	60.46	61.73
MMS	0.68	0.68	-	-	71.50	71.53	-	-	-	-
Stable Diffusion v2.1	0.42	0.43	0.44	0.44	-	-	-	-	-	-

Table 5.4: Evaluation of the visual output quality on the validation and test splits for each modality-specific dataset (PENS for *text*→*text*, MLASK for *text+video*→*text+image*, and M3LS for *text+images*→*text+image*). The highest-scoring system in each column is bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2024).

Visual Output

The relatively high CosSim scores of the random visual baseline (Table 5.4) may indicate that the CLIP features are not distinctive enough for the closely re-

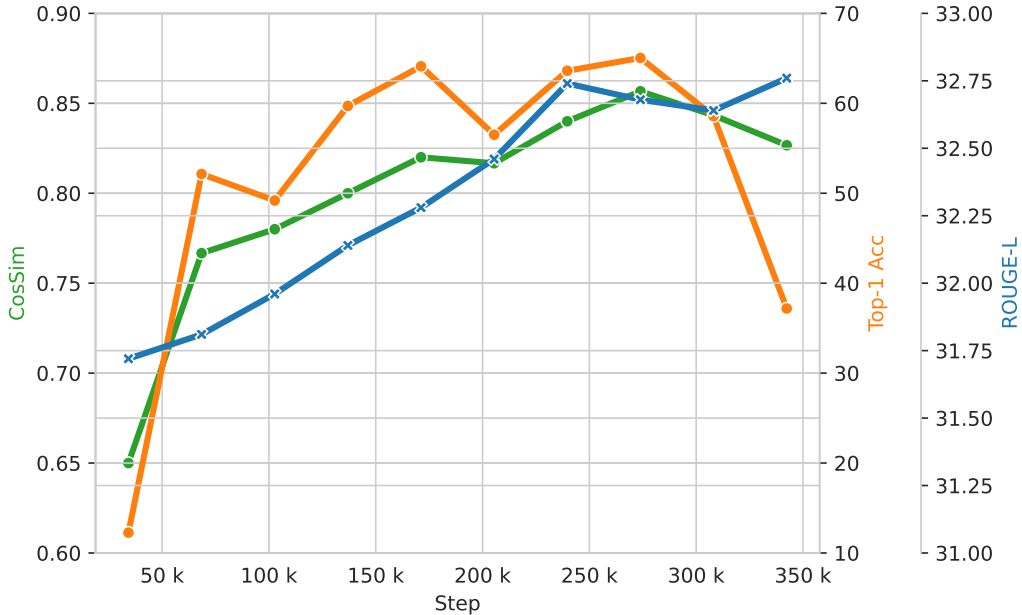


Figure 5.6: The quality of the visual and textual output during the BLIP-2 fine-tuning, reported on the validation split of the M3LS dataset. We report CosSim and Top-1 Acc metrics for the visual output and ROUGE-L for the textual one.

lated images/frames coming from the same article. A similar phenomenon was reported by Fu et al. (2021) (see Section 2.2), who report CosSim of 0.692 with their best model vs CosSim of 0.677 with the random baseline, calculating the similarity based on ResNet-50 features. The model (T5CLIP_{M3LS}) trained solely on the image-based data (M3LS) performs slightly better than the multi-task one (T5CLIP). We attribute this to the potentially easier image-based task formulation (see Section 2.2) where the target input (i.e., one with CosSim = 1.0) is present in the input. The correlation between CosSim and Top-1 Acc is evident – the best-performing T5CLIP_{M3LS} model achieves almost 95% of Accuracy, which evaluates to an average CosSim (0.97) close to 1.0. The Top-1 Acc of the random baseline evaluates to the expected value of roughly 33% – on average, there are roughly three input images (see Section 4.3).

The highest scores of both CLIPBERTScore and CosSim are achieved on MLASK by the MMS model, which uses a separate visual encoder and frame-scoring module. The difference in terms of CosSim is roughly the same between the MMS model and the T5CLIP, as between the T5CLIP and the random baseline. Those results indicate that while the simplified, unified architecture can approach solving the MSMO task, there is still a gap left compared to the sophisticated task-specific architecture. Our solution based on the BLIP-2 model performs noticeably worse than all of the T5CLIP variants. The high CLIPBERTScore scores can be explained by the great quality of textual output – 75% of the CLIPBERTScore is computed based on the BERTScore precision of the input article and the generated summary. The significant drop in Top-1 Acc compared to the T5CLIP variants (~90% → ~60%) can be attributed to the lack of explicit index tokens that force the model to predict a specific string.

By analyzing the scores on the M3LS dev-set during BLIP-2 fine-tuning (see Figure 5.6), we can notice that later into the training, both metrics drop, sug-

gesting a possible over-fitting or slight divergence. On the contrary, the quality of the textual output, as measured with ROUGE-L, constantly increases during the whole training. Surprisingly, despite the drop in the Top-1 Acc by roughly 30 p.p., the value of CosSim decreases by a mere 0.05. Those results indicate a high degree of similarity between input images – picking an image different from the reference does not hurt the average similarity. The visual baseline based on Stable Diffusion scores noticeably below other models. Despite those results, considering the unclear nature of CosSim as an evaluation metric³⁷ and the visual quality and adequacy of the generated cover pictures (see Appendix B.2.2), we believe that the generative approaches to visual summarization will be a prominent research direction in the upcoming years.

5.2.4 Ablation Studies

We designed several auxiliary experiments to provide a deeper look into the unified architecture’s performance.

Firstly, we examine the performance with respect to the frame sampling of the video-based input (see Section 2.2). In our core experiments discussed above, we pre-process the video by sampling the frames at 1fps (on average, 86 frames per video). Considering that the longest video from the MLASK dataset lasts five minutes without any further processing, the unified model would need to consume representations of 300 frames. Such a number of frames is too large to process with the BLIP-2 model – it uses the Q-Former to map each input image into 32 visual tokens, which would require us to process sequences of length up to 9,600. Therefore, we decided to further down-sample the video by sampling 20 frames evenly spaced across the video (on average, 19.9 frames per video – some of the videos do not last 20 seconds). In their foundational work on VMSMO, Li et al. (2020d) considered only 10 frame candidates, but their input videos were shorter, with an average length of one minute.

To examine how this affects the model performance, we trained a variant of T5CLIP_{MLASK} (T5CLIP_{MLASK ALL}) that uses the denser (1 frame per second) sampling for each video. The results on the MLASK dev-set (compared with the T5CLIP_{MLASK} model that samples up to 20 frames) are as follows: ROUGE-L: 20.79 → 20.55, BERTScore: 25.46 → 25.12, CosSim: 0.64 → 0.61 – with more input frames, performance degradation is observed both in terms of textual and visual output. From the textual perspective, those results suggest that the visual clues are not effectively utilized, with more frames introducing mostly noise. From the visual perspective, the value of CosSim may indicate a performance comparable to a random baseline. However, by analyzing the distribution of predicted frames and looking at the training curves, there is a clear learning pattern visible. In our previous experiments with the MLASK dataset (see Section 5.1), we discussed the similarity patterns between the input frames and the target image. By allowing more frames in the input, we potentially introduce more *similar* frames and, thus, dilute the learning signal.

³⁷Since the visual outputs are generated from scratch, the space of image embeddings as computed with feature extractors is not well examined. Even less can be said about the distribution of CosSim scores, computed on such feature vectors.

To further examine this phenomenon, we experimented with a frame sampling algorithm inspired by previous works on detecting scene/shot changes in video – if two frames are similar enough (similarity above a certain threshold), then the expected delta in performance, if we use one instead of the other, is marginal. That allows us to drop one of them, reducing the number of frames. In our implementation, to compare with the uniform sampling, we want to limit ourselves to (up to) 20 frames per video. Therefore, we iteratively remove similar frames from the video until there are no more than 20 frames left. Formally, the algorithm proceeds as follows:

- [1] We assume to be given a sequence of frames $V = (v_1, \dots, v_m)$, the initial threshold τ (0.95), and the maximal desired number of frames Λ (20). We initialize the output (candidate) subsequence $\hat{V} = V$. We proceed by using an off-the-shelf feature extractor to turn the sequence of frames V into a sequence of frame features F (\hat{F}). We iteratively repeat steps [2] and [3], modifying \hat{V} , until we end up with a subsequence $\hat{V} = (v_{i_1}, \dots, v_{i_k})$ of V , such that $k \leq \Lambda$.
- [2] If $k = |\hat{V}| \leq \Lambda$, we stop the algorithm and return \hat{V} . Otherwise, we sample Λ (seed) frames V_{seed} evenly spaced across the candidate frames \hat{V} . Then, we compute the CosSim matrix M between the candidate and the seed features, i.e., $M = \text{CosSim}(\hat{F}, F_{seed}) \in \mathbb{R}^{k \times \Lambda}$. We will say that v_i and v_j are at least τ -close, if $M_{ij} \geq \tau$. Next, we identify frames (rows) i that satisfy $\min_j M_{ij} < \tau$, i.e., frames that are not at least τ -close to any of the seed frames in V_{seed} . If there are no such frames, it means that every candidate frame $v \in \hat{V}$ is at least τ -close to at least one of the frames in V_{seed} . Therefore, we decrease the threshold (in our implementation by 0.05) and jump to step [2]. Otherwise, we jump to step [3].
- [3] We update the collection of candidate frames \hat{V} by dropping those that are at least τ -close to at least one of the frames in V_{seed} , merging the remaining ones with V_{seed} , i.e., $\hat{V} = V_{seed} \cup \{v_i \in \hat{V} : \min_j M_{ij} < \tau\}$, and jump to step [2].

The algorithm is guaranteed to converge as we reduce the number of candidate frames in step [3]. By reducing the threshold τ in step [2], we guarantee that after a finite number of steps, we will identify a frame that is “not at least τ close”, and thus jump to step [3]. If the inverse was true, the set of candidate frames would need to be equal to the set of seed frames and consist only of a single element³⁸. To prove it, it is enough to consider the case of $\hat{V} = V_{seed} \cup \{\hat{v}\}$, i.e., $M = \text{CosSim}(\hat{F}, F_{seed}) \in \mathbb{R}^{1 \times \Lambda}$, the rest follows by induction. Let us assume that the inverse is true, i.e., that $\forall 1 > \varepsilon > 0$, we have $\min_j M_j \geq \varepsilon$. Since $0 \leq \text{CosSim} \leq 1$ and closed intervals in \mathbb{R} are complete, we get that $\min_j M_j = 1$, and thus $\forall j \in \{1, \dots, \Lambda\}; M_j = 1$. But this implies that $\forall v_{seed} \in V_{seed}; v_{seed} = \hat{v}$, and thus $V_{seed} = \{\hat{v}\}$ – contradiction.

After processing all of the videos with the algorithm described above, we trained the T5CLIP_{MLASK ALG} variant with, on average, 19.8 input frames per video. Comparing with the T5CLIP_{MLASK} model, the results on the MLASK dev-set are as follows: ROUGE-L: 20.79 \rightarrow 20.85, BERTScore: 25.46 \rightarrow 25.34,

³⁸“The remaining straightforward but tedious mathematical details are left as an exercise for the interested reader.”

CosSim: 0.64 \rightarrow 0.59. Mixed results are observed, with a marginal improvement in ROUGE-L and a comparable decrease in BERTScore. The value of CosSim is lower than the one achieved by the random baseline despite both the distribution of output labels and the shape of the learning curve not indicating any issues.

Our conclusions are as follows: the frame sampling process can have a non-trivial influence on the quality of both textual and visual output. However, the fact that the textual input is sufficient to generate a high-quality summary means that the models are not greatly affected by the quality of the visual input, putting our observations in line with the “greedy learning” hypothesis by Wu et al. (2022). With respect to the quality of the visual output, we conclude that the CosSim metric is not a clear indicator of the performance and that further research into the multimodal metrics is required (see Section 3.2 and Section 3.3). Our observations are further confirmed by the experiment in which we mask the values of the visual features with random noise and do the inference with the T5CLIP model trained on genuine data. While the Top-1 Acc drops to a chance level (M3LS test-set 88.55 \rightarrow 37.9), the quality (measured with ROUGE-L) of the textual output is not greatly affected (M3LS test 29.38 \rightarrow 29.32).

Secondly, we look at the training process from the perspective of the index tokens. The idea here is to alter the learning signal by focusing more on the visual output, i.e., modifying the loss computation with respect to the index tokens. Those experiments are conducted only with the T5CLIP variant and are enabled by the explicit addition of index tokens to the vocabulary. We explore the smooth labels (see Section 5.1) applied uniformly to both video-based and image-based input³⁹ and a second approach that assigns greater weights w to the visual tokens during Cross-Entropy loss computation⁴⁰. Using 10 times greater weight (T5CLIP_{w=10}) improves the Top-1 Acc on M3LS (compared with T5CLIP, still lags behind the task-specific T5CLIP_{M3LS}), while using 50 times greater weight (T5CLIP_{w=50}) brings no further improvement, degrading the quality of textual output. The smooth labels (T5CLIP_{Smooth}), designed for the video-based data, are not effective on image-based data, as indicated by the low values of both CosSim and Top-1 Acc. Similarly to our experiments on the original MLASK dataset (see Section 5.1), the smooth labels have a minor, positive impact on the quality of the textual output that we attribute to more stabilized training, corresponding to the more wide-spread, general-purpose label smoothing technique (Szegedy et al., 2016).

Thirdly, we wish to examine the model performance on a task-specific dataset that was not explored for training. For that purpose, we explore the test-split of the MSMO dataset (see Section 4.1). This dataset consists of 10,261 English news articles with a varying number of input images. For each article, between 0

³⁹The smooth labels are implemented by modifying the target distribution over all tokens, i.e., we assign the (normalized) similarity to each index token representing an input image/frame, and 0 to all remaining index tokens (the number of input images/frames varies for each document) and the tokens from the original vocabulary.

⁴⁰The weight assigned to the original, non-index tokens is always 1.0. When training with the classical (non-smooth) formulation, the target is a single visual token, and thus, all of the other visual tokens are not affected by the weight, see <https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html>.

	ROUGE-L	BERTScore	% Relevant	% Img Predicted
Alpaca	23.56	26.77	-	-
BRIO	18.64	17.34	-	-
T5CLIP	25.21	27.42	56.2%	95.1%
T5CLIP _{M3LS}	20.79	21.66	58.6%	100.0%
T5CLIP _{w=10}	25.55	27.30	55.3%	61.7%
T5CLIP _{w=50}	26.08	27.40	55.9%	38.5%
T5CLIP _{Smooth}	24.32	26.56	55.7%	91.7%

Table 5.5: Evaluation of the visual and textual output quality, reported on the post-processed test-split of the MSMO dataset (see Section 4.1).

and 6 of the input images were marked as “relevant” by the human annotators. In the first step, we filter the test-data by dropping articles with more than 20 input images and those for which annotators did not mark even a single input image, leaving us with 9,460 instances (92%). In the next step, we drop the articles for which all of the input images were marked as relevant by the annotators, leaving us with 5,063 documents (49%). Since the proposed unified models are trained to predict a single image to act as a pictorial summary, the evaluation must be adapted to the case of multiple image targets. Instead of reporting a simple Top-1 Acc, we will report the percentage of predictions for which the hypothesis belongs to the set of relevant images. By dropping the documents with all of the input images marked as relevant, we make the task non-trivial. Instead of reporting a random baseline, we provide the proportion (across the down-sampled test-set) of input images that are relevant, which corresponds to a trivial baseline that marks all input images and is equal to 50.74%.

The evaluation results are reported in Table 5.5. Besides the ROUGE-L and BERTScore metrics, we report the percentage of predictions belonging to the relevant images and the percentage of instances for which the index token was predicted in the output⁴¹. If the index token is not predicted, we consider the first image as the model output. Models trained in the multi-task fashion seem to out-perform the one trained only on the M3LS dataset (T5CLIP_{M3LS}), suggesting the benefits of additional textual data and the multi-task setup. However, this variant scores the highest in terms of the relevance of the predicted image. The T5CLIP_{w=50} variant predicts the index token only for 39% of articles but performs on par in terms of relevance. This is caused by our fallback formulation – on this test-set, always predicting the first image achieves a relevance score of 55.1%. By comparing with the off-the-shelf textual baselines (Alpaca and BRIO), we prove the validity of our setup, which in the multi-task settings outperforms both of them.

⁴¹When evaluating on the test-splits of the datasets used for fine-tuning, the value was always 100%. Thus, we have not explicitly reported it.

Conclusions

Considering the relative novelty of the task of Multimodal Summarization, our main research goal was not to solve a well-defined problem but rather to provide more insights into the task-specific phenomena and establish foundations for future works.

By curating and releasing the MLASK dataset (see Section 4.2), we created a publicly available artifact that, thanks to its completeness (all of the data modalities are accessible directly via a dedicated data repository, with a transparent training/validation/test split), will hopefully establish itself as one of the standardized benchmarks. The corresponding code-base, thanks to its modular structure and replicable nature, was already used and modified by other researchers in follow-up works (see Section 5.1.4).

Our research targeting the evaluation of the textual output provided yet another link between Machine Translation and Text Summarization. In our work on the COMES metric (see Section 3.1.7), we empirically proved that trainable, estimator-based metrics trained only on (multilingual) MT data are also applicable for evaluating (monolingual) textual summaries. Although less specific than metrics dedicated to summarization, their performance in evaluating the *overall* quality was on par with dedicated solutions. In a symmetrical manner, in our work on the MTEQA metric (see Appendix A), we demonstrated that a content-based evaluation of MT, originally applied mostly to summarization, can be an effective tool for domain-specific texts. Due to the implicit assumptions of the fluency and grammatical correctness of the machine-generated text, the proposed evaluation through the lenses of QA was, by design, not meant as a sole metric to judge the quality of MT systems.

Our experiments with neural networks trained end-to-end for Multimodal Summarization (see Chapter 5) experimentally confirmed a number of our hypotheses but also raised additional questions. We observed that both pre-training (see Section 5.1.3) and multi-task training (see Section 5.2.3) are effective ways of using the high-resource text-only data to improve the quality of the textual output within the MSMO formulation of Multimodal Summarization. Unfortunately, those improvements were generally not accompanied by proportional improvements in terms of the quality of the visual output. An approach that considers the whole distribution of similarities between the input images (frames) and the reference one to formulate the training signal gives some improvement for the video input (see Section 5.1.3). However, those improvements do not carry to an input with multiple images (see Section 5.2.4). It is true that within the image-based formulation of MSMO, the target image is among the input ones, and within the video-based formulation, we only assume the presence of a very similar frame. Nonetheless, the fact that the target images are judged by humans only as “partly relevant” (see Section 5.1.3) undermines the foundations of such analysis. Still, we believe that those issues require further research, as the metrics currently used to evaluate visual outputs (see Section 3.2 and Section 3.3) give us inconsistent and counter-intuitive results (see Section 5.2.4). It must be noted that the VMSMO formulation, which summarizes a video to a single image (frame), may be considered ill-defined. Due to the size difference between the

input (video lasting for a couple of minutes) and the output (a single image), the capacity to carry meaningful (and comparable) information is vastly contrasting.

The unified formulation of **MSMO** (see Section 5.2) that realizes $text+video \rightarrow text+image$, $text+images \rightarrow text+image$, and $text \rightarrow text$ variants of Multimodal Summarization with a single sequence-to-sequence model follows the recent trends of unifying text and vision via Transformer-based models. While the performance generally lags behind the sophisticated, modular architecture, the idea of using visual tokens instead of a dedicated decoder (scoring module) allows for re-formulating the task and applying the **SOTA** decoder-only, vision-capable **LLMs**.

Limitations

It is crucial to recognize that our work and our findings have certain limitations. Some of them are technical and related to the experimental nature of our research.

- When automatically curating the MLASK dataset (see Section 4.2) and collecting the extension of the M3LS dataset (see Section 4.3), a number of algorithmic checks were conducted to assure the quality and adequacy of each data instance. However, a human validation was conducted only on a sample of roughly 100 articles, and thus, it is possible in principle that some corrupted or miss-matched data made it into the final datasets.
- Only a small number of trainings was conducted for each model/variant considered. Thus, our comparisons are mostly point-wise, without considering the error estimates properly.
- Our experiments were enabled by modern, powerful GPUs. The cost related to purchasing and operating such a piece of equipment may reduce the reproducibility of our work, while the necessary energy consumption can have a non-trivial impact on the environment.

The other limitations concern the design part of the presented experiments.

- Our experiments are conducted only on a small number of datasets, including the ones created by ourselves. Since the MLASK dataset is in the Czech language, our findings in Section 5.1 are limited to a single, minor language. As a consequence, when working with English data in Section 5.2, we had to rely on an automatic machine translation system to align the MLASK dataset with other resources. Our choices are due to data sparsity, as discussed in Chapter 4.
- Accordingly, a broader comparison with other architectures/models would be beneficial. While we compared with a number of text-only summarization models, due to the lack of public code-bases and reference checkpoints (model weights) for Multimodal Summarization, we were not able to appropriately compare with other works. The lack of sufficient computational power made it unfeasible to compare with different text-only pre-trained models used to initialize the textual component.
- When exploring the unified formulation of **MSMO** (see Section 5.2), certain variants (see Chapter 2), e.g., $text+video \rightarrow text$, $images \rightarrow text$ or $video \rightarrow text+images$ were not considered.
- In Section 5.1, we took a sequential approach to ablation studies by adding (or removing) components one by one. While a full-scale grid search would

enable better estimation of the influence of each component, it was not feasible with our resources.

Other limitations concern, e.g., the usage of the off-the-shelf generative models (see Section 5.2), which were trained on (partially undisclosed) data that potentially may include harmful content such as sexually explicit materials or toxic, stereotyped language. We did not apply any filtering to the model outputs, so the predictions may not be free of bias. With the field advancing to more practical applications, this could pose a serious threat.

We acknowledge that despite our best efforts, mistakes of human origin (e.g., code-related) may have occurred.

Future Work

Considering how fast-paced the current research on Artificial Intelligence (AI), Machine Learning, Natural Language Processing, and Computer Vision is, we consider making any predictions about the future of Multimodal Summarization a courageous choice. Instead, we rather highlight what we consider major obstacles and challenges that still remain.

- Lack of publicly available datasets and baselines (models) that would be applicable to a variety of data formats (e.g., to both *text+video* and *text+image* inputs).
- Lack of task-specific metrics. To enable proper research on evaluation methods, not only datasets and models but also collections of human annotations are required. Automatic metrics are crucial, as without them, it is not possible to correctly measure the progress of the field.
- The (mostly) extractive nature of the visual component. With the current progress in generative AI, we believe that enabling formulations with text-only input and multimodal output, e.e.,g *text*→*text+image* and *text*→*text+video* would bring the task to a more realistic settings and enable wide-spread practical applications.

Bibliography

- Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. [PolyglotNER: Massive multilingual named entity recognition](#). In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.
- Kamal Al-Sabahi, Zuping Zhang, and Mohammed Nadher. 2018. [A Hierarchical Structured Self-Attentive Model for Extractive Document Summarization](#). *IEEE Access*, 6:24205–24212.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. [Flamingo: a Visual Language Model for Few-Shot Learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.
- Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. [PENS: A dataset and generic framework for personalized news headline generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.
- Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. [Video Summarization Using Deep Neural Networks: A Survey](#). *Proc. IEEE*, 109(11):1838–1863.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. [Neural Machine Translation by Jointly Learning to align and translate](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Yujia Bao, Ankit Parag Shah, Neeru Narang, Jonathan Rivers, Rajeev Maksey, Lan Guan, Louise N. Barrere, Shelley Evenson, Rahul Basole, Connie Miao, Ankit Mehta, Fabien Boulay, Su Min Park, Natalie E. Pearson, Eldhose Joy,

- Tiger He, Sumiran Thakur, Koustav Ghosal, Josh On, Phoebe Morrison, Tim Major, Eva Siqi Wang, Gina Escobar, Jiaheng Wei, Tharindu Cyril Weerasooriya, Queena Song, Daria Lashkevich, Clare Chen, Gyuhak Kim, Dengpan Yin, Don Hejna, Mo Nomeli, and Wei Wei. 2024. [Harnessing Business and Media Insights with Large Language Models](#). *arXiv preprint arXiv:2406.06559*.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020a. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. [Proceedings of the Sixth Conference on Machine Translation](#). Association for Computational Linguistics, Online.
- Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors. 2020b. [Proceedings of the Fifth Conference on Machine Translation](#). Association for Computational Linguistics, Online.
- Araly Barrera and Rakesh M. Verma. 2012. [Combining Syntax and Semantics for Automatic Extractive Single-Document Summarization](#). In *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*, volume 7182 of *Lecture Notes in Computer Science*, pages 366–377. Springer.
- bbc.com. <https://www.bbc.com/>. [Online; accessed May 2024].
- Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. [Re-evaluating evaluation in text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.
- Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. [Prompted opinion summarization with GPT-3.5](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. [CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.
- Spencer Braun, Oleg Vasilyev, Neslihan Iskender, and John Bohannon. 2022. [Does summary evaluation survive translation to other languages?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2425–2435, Seattle, United States. Association for Computational Linguistics.
- Sergey Brin and Lawrence Page. 1998. [The anatomy of a large-scale hypertextual web search engine](#). *Computer Networks and ISDN Systems*, 30(1):107–117. Proceedings of the Seventh International World Wide Web Conference.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language Models are Few-Shot Learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. [Further meta-evaluation of machine translation](#). In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.
- Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal Sentence Encoder](#). *arXiv preprint arXiv:1803.11175*.
- Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022. [VLP: A Survey on Vision-language Pre-training](#). *Machine Intelligence Research*, 20:38–56.
- Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. [Towards knowledge-based personalized product description generation in e-commerce](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, page 3040–3050, New York, NY, USA. Association for Computing Machinery.
- Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. [Extending Context Window of Large Language Models via Positional Interpolation](#).

- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. [Unifying Vision-and-Language Tasks via Text Generation](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2023. [Scaling Instruction-Finetuned Language Models](#). *arXiv preprint arXiv:2210.11416*.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. [Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling](#). *arXiv preprint arXiv:1412.3555*.
- Jacob Cohen. 1960. [A Coefficient of Agreement for Nominal Scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Pierre Jean A. Colombo, Chloé Clavel, and Pablo Piantanida. 2022. [Infolm: A new metric to evaluate summarization & data2text generation](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10554–10562. AAAI Press.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- daily mail.co.uk. www.dailymail.co.uk/sciencetech/article-10995713/NA_SAs-James-Webb-Telescope-targets-Carina-Nebula-Stephans-Quintet-Southern-Ring-Nebula-more.html. [Online; accessed July 2022].
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [ImageNet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021a. [Towards question-answering as an automatic metric for evaluating the content quality of a summary](#). *Transactions of the Association for Computational Linguistics*, 9:774–789.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

- Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. [Re-examining system-level correlations of automatic summarization evaluation metrics](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.
- Daniel Deutsch, George Foster, and Markus Freitag. 2023. [Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.
- Daniel Deutsch and Dan Roth. 2022. [Benchmarking answer verification methods for question answering-based summarization evaluation metrics](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3759–3765, Dublin, Ireland. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Esin Durmus, He He, and Mona Diab. 2020. [FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.
- H. P. Edmundson. 1969. [New Methods in Automatic Extracting](#). *J. ACM*, 16:264–285.
- Ori Bar El, Ori Licht, and Netanel Yosephian. 2019. [Gilt: Generating images from long text](#). *arXiv preprint arXiv:1901.02404*.
- Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2020. [Edgesumm: Graph-based framework for automatic text summarization](#). *Information Processing and Management*, 57(6):102264.
- Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. [Automatic text summarization: A comprehensive survey](#). *Expert Systems with Applications*, 165:113679.

- Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. [Generating medical reports from patient-doctor conversations using sequence-to-sequence models](#). In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.
- Günes Erkan and Dragomir R. Radev. 2004. [Lexrank: Graph-based centrality as salience in text summarization](#). *Journal of Artificial Intelligence Research*.
- Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. [Question answering as an automatic evaluation metric for news article summarization](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Ali Faheem, Faizad Ullah, Muhammad Sohaib Ayub, and Asim Karim. 2024. [UrduMASD: A multimodal abstractive summarization dataset for Urdu](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17245–17253, Torino, Italy. ELRA and ICCL.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. [Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [GPTScore: Evaluate as You Desire](#). *arXiv preprint arXiv:2302.04166*.
- Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. [Multi-modal Summarization for Video-containing Documents](#). *arXiv preprint arXiv:2009.08018*.
- Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. [MM-AVS: A full-scale dataset for multi-modal summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, Online. Association for Computational Linguistics.
- Mingqi Gao, Wenqing Wang, Xiaojun Wan, and Yuemei Xu. 2023. [Evaluating factuality in cross-lingual summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12415–12431, Toronto, Canada. Association for Computational Linguistics.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. [Large-scale weakly-supervised pre-training for video action recognition](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12038–12047.
- Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. [COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. [Generative Adversarial Nets](#). In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.
- Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. [Accurate evaluation of segment-level machine translation metrics](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.
- Max Grusky. 2023. [Rogue scores](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.

- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. [Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.
- Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. [Vision-and-language navigation: A survey of tasks, methods, and future directions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. [xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection](#). *arXiv preprint arXiv:2310.10482*.
- HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. [Bridging background knowledge gaps in translation with automatic explicitation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.
- HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. 2022. [SimQA: Detecting simultaneous MT errors through word-by-word question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. [Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?](#) In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555.
- Donna Harman and Paul Over. 2002. [The DUC summarization evaluations](#). In *Proceedings of the Second International Conference on Human Language Technology Research, HLT '02*, page 44–51, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XLsum: Large-scale multilingual abstractive summarization for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep Residual Learning for Image Recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In

- Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Wang. 2023. [Let’s think frame by frame with VIP: A video infilling and prediction dataset for evaluating video chain-of-thought](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 204–219, Singapore. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Comput.*, 9(8):1735–1780.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. [TRUE: Re-evaluating factual consistency evaluation](#). In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.
- Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. [MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications](#). *arXiv preprint arXiv:1704.04861*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.
- Lifeng Hua, Xiaojun Wan, and Lei Li. 2018. [Overview of the NLPCC 2017 Shared Task: Single Document Summarization](#). In *Natural Language Processing and Chinese Computing*, pages 942–947, Cham. Springer International Publishing.
- Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. [What have we achieved on text summarization?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.
- Kung-Hsiang Huang, Philippe Laban, Alexander R Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. [Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles](#). *arXiv preprint arXiv:2309.09369*.
- Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. [Self-supervised multimodal opinion summarization](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

- Long Papers*), pages 388–403, Online. Association for Computational Linguistics.
- Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. [A Survey on Multi-modal Summarization](#). *ACM Comput. Surv.*, 55(13s).
- Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. [3D Convolutional Neural Networks for Human Action Recognition](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.
- Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. [Exploiting pseudo image captions for multimodal summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 161–175, Toronto, Canada. Association for Computational Linguistics.
- Liqliang Jing, Jingxuan Zuo, and Yue Zhang. 2024. [Fine-grained and Explainable Factuality Evaluation for Multimodal Summarization](#). *arXiv preprint arXiv:2402.11414*.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Trans. Assoc. Comput. Linguistics*, 5:339–351.
- Jeeweon Jung, Roshan Sharma, William Chen, Bhiksha Raj, and Shinji Watanabe. 2024. [AugSumm: towards generalizable speech summarization using synthetic labels from large language model](#). *arXiv preprint arXiv:2401.06806*.
- Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. [MetricX-23: The Google submission to the WMT 2023 metrics shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. [The Kinetics Human Action Video Dataset](#). *arXiv preprint arXiv:1705.06950*.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. [CTRL - A Conditional Transformer Language Model for Controllable Generation](#). *arXiv preprint arXiv:1909.05858*.
- Aman Khullar and Udit Arora. 2020. [MAST: Multimodal abstractive summarization with trimodal hierarchical attention](#). In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69, Online. Association for Computational Linguistics.

- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. [ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A Method for Stochastic Optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. [To ship or not to ship: An extensive evaluation of automatic metrics for machine translation](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. [Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies](#). *arXiv preprint arXiv:2401.06760*.
- Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névoul, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors. 2022. *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).
- Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors. 2023. *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.
- Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. [FFCI: A Framework for Interpretable Automatic Evaluation of Summarization](#). *J. Artif. Int. Res.*, 73.
- Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. [Evaluating the efficacy of summarization evaluation across languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.
- Mahnaz Koupaee and William Yang Wang. 2018. [WikiHow: A Large Scale Text Summarization Dataset](#). *arXiv preprint arXiv:1810.09305*.

- Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021a. [Just ask! evaluating machine translation by asking and answering questions](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.
- Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021b. [MTEQA at WMT21 metrics shared task](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1024–1029, Online. Association for Computational Linguistics.
- Mateusz Krubiński and Pavel Pecina. 2022. [From COMET to COMES – can summary evaluation benefit from translation evaluation?](#) In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 21–31, Online. Association for Computational Linguistics.
- Mateusz Krubiński and Pavel Pecina. 2023. [MLASK: Multimodal summarization of video-based news articles](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 910–924, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mateusz Krubiński and Pavel Pecina. 2024. [Towards unified uni- and multi-modal news headline generation](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 437–450, St. Julian’s, Malta. Association for Computational Linguistics.
- Mateusz Krubiński. 2022. [Multimodal Summarization](#). *ÚFAL MFF UK Ph.D. Thesis Proposal*.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

- Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. [Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.
- Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. [Multi-modal sentence summarization with modality attention and image filtering](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.
- Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. [Multi-modal summarization for asynchronous collection of text, image, audio and video](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. [Multimodal sentence summarization via multimodal selective encoding](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020c. [HERO: Hierarchical encoder for Video+Language omni-representation pre-training](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.
- Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020d. [VMSMO: Learning to generate multimodal summary for video-based news articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. [Leveraging Large Language Models for NLG Evaluation: A Survey](#). *arXiv preprint arXiv:2401.07103*.

- Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2023a. [D²TV: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14910–14922, Singapore. Association for Computational Linguistics.
- Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023b. [Summary-oriented vision modeling for multimodal abstractive summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2951, Toronto, Canada. Association for Computational Linguistics.
- Jindřich Libovický and Jindřich Helcl. 2017. [Attention strategies for multi-source sequence-to-sequence learning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.
- Rensis Likert. 1932. [A technique for the measurement of attitudes](#). *Archives of Psychology*, 22(140):55–55.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chin-Yew Lin and Eduard Hovy. 2003. [Automatic evaluation of summaries using n-gram co-occurrence statistics](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.
- Dengtian Lin, Liqiang Jing, Xuemeng Song, Meng Liu, Teng Sun, and Liqiang Nie. 2023a. [Adapting Generative Pretrained Language Model for Open-domain Multimodal Sentence Summarization](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 195–204, New York, NY, USA. Association for Computing Machinery.
- Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023b. [VideoXum: Cross-modal Visual and Textural Summarization of Videos](#). In *IEEE Transactions on Multimedia*, pages 1–13. IEEE.
- Wuhang Lin, Shasha Li, Chen Zhang, Bin Ji, Jie Yu, Jun Ma, and Zibo Yi. 2022. [SummScore: A Comprehensive Evaluation Metric for Summary Quality Based on Cross-Encoder](#).
- Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. [Multistage fusion with forget gate for multimodal summarization in open-domain videos](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845, Online. Association for Computational Linguistics.

- Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022a. [Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6959–6969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019a. [Hierarchical transformers for multi-document summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019b. [Text summarization with pretrained encoders](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *CoRR*, abs/1907.11692.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. [Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. [BRIO: Bringing order to abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.
- Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023c. [On Learning to Summarize with Large Language Models as References](#). *arXiv preprint arXiv:2305.14239*.
- Yu Lu Liu, Rachel Bawden, Thomas Scaliom, Benoît Sagot, and Jackie Chi Kit Cheung. 2022c. [MaskEval: Weighted MLM-Based Evaluation for Text Summarization and Simplification](#). *CoRR*, abs/2205.12394.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

- Alejandro Lozano, Scott L. Fleming, Chia-Chun Chiang, and Nigam Shah. 2024. [Climfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature](#). *Biocomputing 2024*, pages 8–23.
- Hans Peter Luhn. 1958. [The Automatic Creation of Literature Abstracts](#). *IBM J. Res. Dev.*, 2:159–165.
- Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. [Multi-document Summarization via Deep Learning Techniques: A Survey](#). *ACM Comput. Surv.*, 55(5).
- Kateřina Macková and Milan Straka. 2020. [Reading Comprehension in Czech via Machine Translation and Cross-Lingual Transfer](#). In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, page 171–179, Berlin, Heidelberg. Springer-Verlag.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali. Association for Computational Linguistics.
- Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. [Results of the WMT20 metrics shared task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. [End-to-End Learning of Visual Representations from Uncurated Instructional Videos](#). In *CVPR*.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. [HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips](#). In *ICCV*.
- Rada Mihalcea and Paul Tarau. 2004. [TextRank: Bringing order into text](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient Estimation of Word Representations in Vector Space](#). In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. [Where’s the point? self-supervised multilingual punctuation-agnostic sentence segmentation.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.
- Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. [Trends in Integration of Vision and Language Research: A Survey of tasks, datasets, and methods.](#) *J. Artif. Intell. Res.*, 71:1183–1317.
- Arie Nakhmani and Allen Tannenbaum. 2013. [A New Distance Measure Based on Generalized Image Normalized Cross-Correlation for Robust Video Tracking and Image Recognition.](#) *Pattern Recognit Lett*, 34(3):315–321.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. [SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents.](#) In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond.](#) In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Ani Nenkova and Rebecca Passonneau. 2004. [Evaluating content selection in summarization: The pyramid method.](#) In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.
- OpenAI. 2024. [GPT-4 Technical Report.](#) *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback.](#)

In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

- Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. [mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore. Association for Computational Linguistics.
- Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. [Multimodal abstractive summarization for how2 videos](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.
- Pinelopi Papalampidi and Mirella Lapata. 2023. [Hierarchical3D adapters for long video-to-text summarization](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320, Dubrovnik, Croatia. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. [Kosmos-2: Grounding Multimodal Large Language Models to the World](#). *arXiv preprint arXiv:2306.14824*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. [Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals](#). *Nature Communications*, 11(4381):1–15.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. [The Kaldi Speech Recognition Toolkit](#). In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. [Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

- Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. [Data-to-text generation with content selection and planning](#). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’19/IAAI’19/EAAI’19. AAAI Press.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Lingfeng Qiao, Chen Wu, Ye Liu, Haoyuan Peng, Di Yin, and Bo Ren. 2022. [Grafting pre-trained models for multimodal headline generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 244–253, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, and Lijuan Wang. 2024. [MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21909–21921.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning Transferable Visual Models From Natural Language Supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language Models are Unsupervised Multitask Learners](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. [Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-Resolution Image Synthesis with Latent Diffusion Models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.
- Huan Rong, Zhongfeng Chen, Zhenyu Lu, Fan Xu, and Victor S. Sheng. 2024. [Multization: Multi-Modal Summarization Enhanced by Multi-Contextually Relevant and Irrelevant Attention Alignment](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [ImageNet Large Scale Visual Recognition Challenge](#). *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barraud, Lucia Specia, and Florian Metze. 2018. [How2: a large-scale dataset for multimodal language understanding](#). In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. [MLSUM: The multilingual summarization corpus](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.
- Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. [Answers unite! unsupervised metrics for reinforced summarization models](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Seznam Zprávy API. <https://api-web.seznamzpravy.cz/v1/documents/213174>. [Online; accessed May 2024].
- seznamzpravy.cz. <https://www.seznamzpravy.cz/clanek/tech-technologie-nasa-podruhe-zrusila-start-rakety-k-mesici-podivejte-se-jak-mela-letet-213174>. [Online; accessed November 2023].
- Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. [Crowdsourcing lightweight pyramids for manual summary evaluation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.
- Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. 2022. [End-to-end speech summarization using restricted self-attention](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076.
- Noam Shazeer and Mitchell Stern. 2018. [Adafactor: Adaptive Learning Rates with Sublinear Memory Cost](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.
- Faisal Tareque Shohan, Mir Tafseer Nayeem, Samsul Islam, Abu Ubaida Akash, and Shafiq Joty. 2024. [XL-HeadTags: Leveraging Multimodal Retrieval Augmentation for the Multilingual Generation of News Headlines and Tags](#). *arXiv preprint arXiv:2406.03776*.
- Karen Simonyan and Andrew Zisserman. 2014. [Very deep convolutional networks for large-scale image recognition](#). *arXiv preprint arXiv:1409.1556*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Milan Straka, Jan Hajič, and Jana Straková. 2016. [UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing](#). In *Proceedings of the Tenth International Conference*

- on *Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. [SumeCzech: Large Czech news-based summarization dataset](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [RoFormer: Enhanced transformer with Rotary Position Embedding](#). *Neurocomputing*, 568:127063.
- Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. 2021. [Non-autoregressive text generation with pre-trained language models](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 234–243, Online. Association for Computational Linguistics.
- Dima Suleiman and Arafat A. Awajan. 2020. [Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges](#). *Mathematical Problems in Engineering*, 2020:1–29.
- Maosong Sun, Jingyang Li, Zhipeng Guo, Zhao Yu, Yabin Zheng, Xiance Si, and Zhiyuan Liu. 2016. [THUCTC: An Efficient Chinese Text Classifier](#).
- Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. 2024. [From Sora What We Can See: A Survey of Text-to-Video Generation](#). *arXiv preprint arXiv:2405.10674*.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. [Rethinking the inception architecture for computer vision](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Mingxing Tan and Quoc Le. 2019. [EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.
- Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. [Effective LSTMs for target-dependent sentiment classification](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

- Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2024. [TLDW: Extreme Multimodal Summarization of News Videos](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1469–1480.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
- Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. [Efficient transformers: A survey](#). *ACM Comput. Surv.*, 55(6):109:1–109:28.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Abhisek Tiwari, Shreyangshu Bera, Sriparna Saha, Pushpak Bhattacharyya, and Samrat Ghosh. 2024. [Yes, This Is What I Was Looking For! Towards Multimodal Medical Consultation Concern Summary Generation](#). In *Advances in Information Retrieval*, pages 210–225, Cham. Springer Nature Switzerland.
- Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. [Multi-modal summarization of key events and top players in sports tournament videos](#). In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 471–478.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [LLaMA: Open and Efficient Foundation Language Models](#). *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madsen Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kam-badur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.

- Ni Trieu, Sebastian Goodman, P. Narayana, Kazuo Sone, and Radu Soricut. 2020. [Multi-Image Summarization: Textual Summary from a Set of Cohesive Images](#). *arXiv preprint arXiv:2006.08686*.
- Dusan Varis and Ondřej Bojar. 2021. [Sequence length is a domain: Length-based overfitting in transformer models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. [Large scale multi-lingual multi-modal summarization dataset](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632, Dubrovnik, Croatia. Association for Computational Linguistics.
- Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. [TL;DR: Mining Reddit to learn automatic summarization](#). In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.
- David Wan and Mohit Bansal. 2022. [Evaluating and improving factuality in multimodal abstractive summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. [Asking and Answering Questions to Evaluate the Factual Consistency of Summaries](#). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. [GIT: A Generative Image-to-text Transformer for Vision and Language](#). *Transactions on Machine Learning Research*.
- Lanxiao Wang, Wenzhe Hu, Heqian Qiu, Chao Shang, Taijin Zhao, Benliu Qiu, King Ngi Ngan, and Hongliang Li. 2022b. [A Survey of Vision and Language Related Multi-Modal Task](#). *CAAI Artificial Intelligence Research*, 1(2):111–136.

- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022c. [OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. [Emergent Abilities of Large Language Models](#). *Trans. Mach. Learn. Res.*, 2022.
- Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. [Review of automatic text summarization techniques & methods](#). *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. [Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 24043–24055. PMLR.
- Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. [CF-Sum coarse-to-fine contribution network for multimodal summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational*

- Linguistics (Volume 1: Long Papers)*, pages 8538–8553, Toronto, Canada. Association for Computational Linguistics.
- Shi Xiaorui. 2023. [MCLS: A large-scale multimodal cross-lingual summarization dataset](#). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 862–874, Harbin, China. Chinese Information Processing Society of China.
- Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. [Factual consistency evaluation for text summarization via counterfactual estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- P. Xu, X. Zhu, and D. A. Clifton. 2023a. [Multimodal Learning With Transformers: A Survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.
- Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023b. [BridgeTower: Building Bridges between Encoders in Vision-Language Representation Learning](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10637–10647. AAAI Press.
- Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023c. [BridgeTower: Building Bridges between Encoders in Vision-Language Representation Learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10637–10647.
- Zenan Xu, Xiaojun Meng, Yasheng Wang, Qinliang Su, Zexuan Qiu, Xin Jiang, and Qun Liu. 2023d. [Learning summary-worthy visual representation for abstractive summarization in video](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. [Automatic Text Summarization Methods: A Comprehensive Review](#). *arXiv preprint arXiv:2204.01849*.
- Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. [Visual goal-step inference using wikiHow](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. [A Survey on Multimodal Large Language Models](#). *arXiv preprint arXiv:2306.13549*.
- Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. [Vision guided generative pre-trained language models for multimodal abstractive summarization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. [Extractive summarization via ChatGPT for faithful summary generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. [PEGA-SUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. [Vision-Language Models for Vision Tasks: A Survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.
- Litian Zhang, Xiaoming Zhang, Ziming Guo, and Zhipeng Liu. 2023b. [CISum: Learning Cross-modality Interaction to Enhance Multimodal Semantic Coverage for Multimodal Summarization](#). In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pages 370–378. SIAM.
- Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022a. [Hierarchical Cross-Modality Semantic Correlation Learning Model for Multimodal Summarization](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11676–11684.
- Mengli Zhang, Gang Zhou, Wanting Yu, Ningbo Huang, and Wenfen Liu. 2022b. [A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning](#). *Computational Intelligence and Neuroscience*, 2022.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. [BERTScore: Evaluating Text Generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022c. [UniMS: A Unified Framework for Multimodal Summarization with Knowledge Distillation](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11757–11764. AAAI Press.

- Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. [Neural document summarization by jointly learning to score and select sentences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.
- Yutong Zhou and Nobutaka Shimada. 2023. [Vision + Language Applications: A Survey](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 826–842.
- Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. [MSMO: Multimodal summarization with multimodal output](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.
- Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. [Multimodal summarization with guidance of multimodal reference](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

A. MTEQA

A.1 Overview

This chapter is based on the JUST ASK! EVALUATING MACHINE TRANSLATION BY ASKING AND ANSWERING QUESTIONS (Krubiński et al., 2021a) and MTEQA AT WMT21 METRICS SHARED TASK (Krubiński et al., 2021b) articles.

In Chapter 3, we briefly introduced the MTEQA (Machine Translation Evaluation with Question Answering) metric for MT evaluation, which we designed inspired by the QA-based approach to summary evaluation. In this Appendix, we will provide an overview of our work.

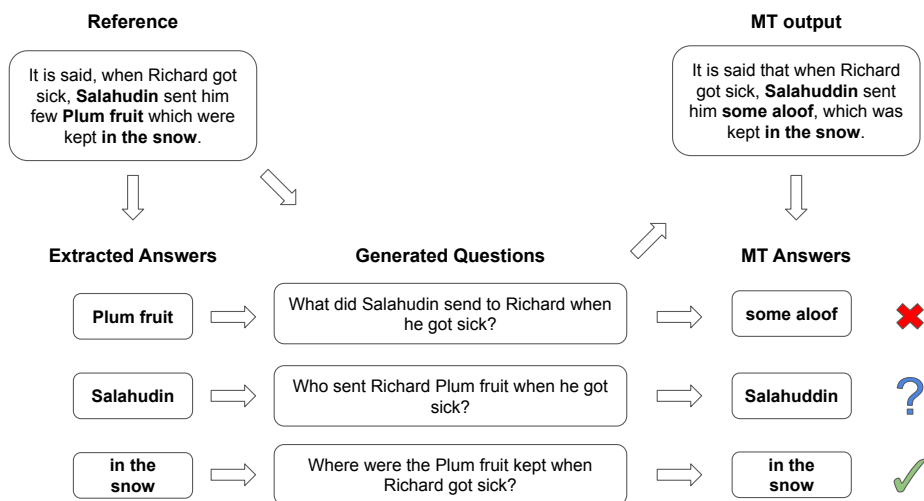


Figure A.1: An illustration of the MTEQA pipeline. One of the MT answers is clearly wrong, one is correct but the other differs with just a single character, raising a question about the choice of the answer-comparison metric. Figure reprint from Krubiński et al. (2021a).

The MTEQA metric builds upon the fact that state-of-the-art (neural) MT systems tend to produce a fluent output but sometimes fail in the adequacy of the translation. Therefore, our goal was to design an automatic metric that could pinpoint the potentially miss-translated phrases and judge to what extent the reference phrasing differs from the one in the MT hypothesis.

To do so, we leverage the automatic QG and QA systems to formulate and answer human-readable questions about the MT system output. Specifically, to check whether a piece of information is preserved, we automatically generate pairs of a question and its (gold-standard) answer from the reference translation and employ a question-answering system to provide a new (test) answer given the question and the MT output (translation) used as the context. The generated

(test) answer is then compared to the gold-standard answer. We assume that if it was possible to answer a question looking only at the reference, it should also be possible to answer this question looking only at the MT output and that the two answers should be identical or very similar. In principle, the proposed MTEQA metric requires solving the following tasks, see Figure A.1:

1. **Answer extraction** – identifies the key information in a sentence, which should also be present in the MT output. This extraction can be treated in a hierarchical/nested manner. For instance, given the sentence “*Today for dinner I had an organic pasta with garlic.*”, the question “*What did you have for dinner today?*” can be correctly answered by all the following phrases *pasta*, *organic pasta* and *organic pasta with garlic*. Thus, answer extraction is performed first, and the questions are generated afterward for each of the answers independently. The same question can be paired with multiple (nested) answers which allows capturing a partial correspondence.
2. **Question generation** – given a reference translation, produces a human-readable question, for which a given keyphrase is the correct answer. For each of the extracted answers, each question is generated independently from the other answers.
3. **Question answering** – generates an answer, given a natural language question and a sentence used as a context. Since we assume that the MT output should carry enough information to answer any question asked based on the reference, we do not consider the non-answerable questions.
4. **Answer comparison** – assesses to what extent the generated answer is correct, given the gold-standard answer extracted from the reference. Metrics based on exact match should be avoided because they are too strict. For example, given the gold-standard answer “*Tchaikovsky*”, both the “*Tchaikovski*” and “*Beethoven*” would get the same score.

Formally, for a given segment s_i , reference translation r_i and MT system output t_i , MTEQA proceeds as follows:

1. Generate the gold-standard answers $a_{i1}, a_{i2}, \dots, a_{ik}$ from the reference r_i
2. For each answer a_{ij} and reference r_i , generate a natural language question q_{ij}
3. Answer each question q_{ij} using the MT output t_i as a context, obtaining answer \tilde{a}_{ij}
4. The final score for a given translation of a segment s_i is the average over all of the generated questions:

$$MTEQA(t_i) = \frac{\sum_1^k D(a_{ij}, \tilde{a}_{ij})}{k},$$

where $D(\cdot, \cdot)$ is a string-comparison metric used to compare the two answers, and k is the number of gold-standard answers extracted from the reference.

A.2 Implementation

In our initial experiments, we have explored the T5 model (Raffel et al., 2020) fine-tuned on the SQuAD dataset (Rajpurkar et al., 2016) for all three sub-tasks, i.e., answer extraction, question generation, and question answering. We

	cs-en	de-en	zh-en		en-de	en-cs
	12	12	16	<i>avg</i>	14	12
MTEQA F1	0.782	0.997	0.952	0.893	0.946	0.845
SENTBLEU	0.844	0.978	0.948	0.859	0.934	0.840
BLEU	0.851	0.985	0.956	0.854	0.928	0.825
PRISM	0.818	0.998	0.957	0.880	0.958	0.949
YISI-2	0.764	0.988	0.964	0.821	0.899	0.714

Table A.1: System-level Pearson correlation for selected metrics used for measuring MT quality with DA human assessment over MT systems using the *newstest2020* references. Average (*avg*) is computed over all to-English directions available. A number below the language pair indicates the number of systems considered. Table reprint from Krubiński et al. (2021a).

Pattern	Extracted Answer	Sentence
NOUN	Coldplay	... the British rock group Coldplay with special guest performers ...
ADJ NOUN	natural grass	As is customary for Super Bowl games played at natural grass stadiums ...
DET NOUN	a fumble	... including a fumble which they recovered for a touchdown ...
NUM NOUN	10 times	The South Florida/Miami area has previously hosted the event 10 times ...
PROPN PROP	Carolina Panthers	... the National Football Conference (NFC) champion Carolina Panthers ...
DET ADJ NOUN	A professional fundraiser	A professional fundraiser will aid in finding business sponsors ...
DET VERB NOUN	a broken arm	... went down with a broken arm in the NFC Championship Game ...
NUM PUNCT NUM	15-1	The Panthers finished the regular season with a 15-1 record ...
DET NOUN ADP NOUN	the application of electricity	Tesla theorized that the application of electricity to the brain ...

Table A.2: Examples of the most frequent POS patterns of gold-standard answers in the XQuAD dataset that we explored to create the POS pattern bank. Table reprint from Krubiński et al. (2021a).

evaluated the metrics based on the submissions to the WMT20 News translation task (Barrault et al., 2020a) and their (direct) human assessments (DA), see Table A.1. We report individual results for selected translation directions into English plus aggregated results (average) for all to-English directions which were part of the WTM20 Metric Task (Barrault et al., 2020a) evaluation campaign¹.

The basic variant of MTEQA used the world-level F1 metric (MTEQA F1), following the classical evaluation protocol for QA. We have compared several lexical similarity metrics (see Table A.3) for the answer comparison step, obtaining the best average results with BLEU (MTEQA BLEU).

Having observed that the baseline model generates, on average, roughly 2 answers per reference – and the same number of questions, as a single question is generated for each answer – we explored additional ways of generating valid questions, as they constitute the whole predictive power of MTEQA. Firstly, we considered exploiting the MT output as an additional source of question/answer pairs. After following the standard procedure, we swapped the roles of MT output and reference – we generated gold-standard answers and questions from the MT output and used reference as a context to answer them. This mimics the approach to precision/recall in, e.g., BERTScore. We did not consider the source segment, as it would have required a cross-lingual QA system. Secondly, we explored external methods to mark keyphrases (potential answers) in the reference. Namely, given a sentence as the input, we parsed it using UDPipe (Straka et al., 2016) to extract part of speech (POS) tags. Then, we extracted phrases match-

¹cs, de, ja, pl, ru, ta, zh, iu, km, ps → en

	cs-en	de-en	zh-en	ja-en	ru-en	ps-en	<i>avg</i>
MTEQA F1	0.782	0.997	0.952	0.982	0.908	0.982	0.893
MTEQA CHRf	0.796	0.996	0.959	0.982	0.901	0.980	0.887
MTEQA BLEU	0.762	0.998	0.954	0.983	0.925	0.985	0.894
MTEQA EXACT	0.762	0.998	0.954	0.966	0.910	0.986	0.883
MTEQA F1 OUT	0.808	0.998	0.949	0.980	0.917	0.984	0.891
MTEQA CHRf OUT	0.835	0.997	0.957	0.979	0.910	0.986	0.891
MTEQA BLEU OUT	0.809	0.998	0.950	0.981	0.929	0.984	0.896
MTEQA EXACT OUT	0.827	0.999	0.948	0.969	0.902	0.983	0.884
MTEQA F1 KP	0.851	0.998	0.944	0.978	0.930	0.986	0.896
MTEQA CHRf KP	0.890	0.998	0.951	0.978	0.927	0.981	0.905
MTEQA BLEU KP	0.844	0.998	0.939	0.973	0.945	0.991	0.900
MTEQA EXACT KP	0.858	0.997	0.938	0.959	0.936	0.990	0.893
MTEQA F1 OUT KP	0.831	0.998	0.942	0.978	0.914	0.992	0.893
MTEQA CHRf OUT KP	0.851	0.998	0.947	0.977	0.917	0.990	0.902
MTEQA BLEU OUT KP	0.842	0.998	0.938	0.971	0.913	0.990	0.895
MTEQA EXACT OUT KP	0.838	0.998	0.936	0.960	0.918	0.992	0.887

Table A.3: System-level Pearson correlation for various variants of the proposed MTEQA metric with DA human assessment over MT systems using the *newstest2020* references. The average is computed over all to-English directions available. Table reprint from Krubiński et al. (2021a).

ing one of the patterns in our POS pattern bank. The POS pattern bank was created by parsing the sentences from XQuAD (Artetxe et al., 2020) dataset, extracting the POS patterns corresponding to the gold-standard answers, and taking the most frequent patterns. This dataset contains professional translations of the development set of SQuAD, translated into various languages from different language families and using different scripts. Table A.2 shows some examples of the extracted POS patterns. Additionally, we extracted named entities mentioned in the reference using a combination of two multilingual NER models, POLYGLOT-NER (Al-Rfou et al., 2015), and Stanza (Qi et al., 2020). The union of the extracted phrases and named entities was considered as the potential answers, see an example in Table A.4. On average, this yielded roughly 7 answers/questions per reference.

When evaluating the metric performance (see Table A.3), we observed that the average correlation with human judgments – the way of measuring metric performance – obtained using the MT output to generate questions (rows with OUT) was very similar but slightly worse than the one using just the questions from the reference. However, the method based on POS pattern matching and NER (rows with KP) yielded improvements over various translation directions. After all, the best-performing configuration of MTEQA was based on answer extraction with POS pattern matching and NER together with the chrF metric used for answer comparison (MTEQA CHRf KP). Our implementation and setup are publicly available at <https://github.com/ufal/MTEQA>.

Answer	Question
<i>Answers extracted using the method based on POS sequences and NER tags</i>	
the stadium	Where did the cat fall from?
an American football match	At what event did spectators catch a cat?
upper deck	What part of the stadium did the cat fall from?
A cat	What animal was caught by spectators at an American football match in Miami Gardens?
Florida	Where is Miami Gardens located?
spectators	Who caught a cat at an American football match in Miami Gardens?
Miami Gardens	Where was a cat caught by spectators at an American football match?
<i>Answers extracted using the baseline model</i>	
cat	What animal was caught by spectators at a football match in Miami Gardens?
Miami Gardens	Where was a cat caught by spectators at an American football match?

Table A.4: Extracted keyphrases and generated corresponding questions for the sentence: “*A cat was caught by spectators at an American football match in Miami Gardens, Florida, after it fell from the stadium’s upper deck.*”, that compare the baseline and keyphrase extraction method based on POS pattern matching and NER. Table reprint from Krubiński et al. (2021a).

A.3 Discussion

To compare our solution against other SOTA metrics, we submitted the MTEQA metric (Krubinski et al., 2021b) to the WMT21 Metric Shared Task. One of the limitations of MTEQA is the requirement for a QA/QG system in the target language. This prohibited us from computing scores for every translation direction/target language. Besides English (Hausa, German, Czech, Russian, Chinese, Japanese, Icelandic→English), we have also submitted results for language pairs with German (English, French→German) and Czech (English→Czech) as target languages. The non-English QA/QG systems were trained by fine-tuning the multilingual mT5 model (Xue et al., 2021) on the machine-translated SQuAD. We exploited the existing translations into German by Lewis et al. (2020b) and into Czech by Macková and Straka (2020). Performance of those systems on *newstest2020* is reported in Table A.1, in columns “en-de” and “en-cs”.

The results were mixed. In the news domain, which was the core of WMT evaluations in recent years, MTEQA was in the middle of the pack for system-level correlations and tended to score below average in the segment-level correlations, see Table A.3 for the Hausa→English direction as an example. However, MTEQA achieved the highest correlation with human annotators on the challenging English→Chinese test-set based on TED talks, see Figure A.2. Those results suggest that while MTEQA should not be used on its own as a sole metric to measure the performance of MT system, it has its applications in niche domains when a specialized, targeted evaluation is essential.

Still, the QA-based approach to MT evaluation that we proposed caught some attention in the scientific community. Han et al. (2022) propose the word-by-word

	ref-A	ref-B
MQM	5.52	0.42
MTEQA	0.47 (3)	0.74 (1)
TER	0.40 (9)	0.71 (2)
BERTScore	0.42 (6)	0.69 (3)
bleurt-20	0.45 (5)	0.68 (4)
cushLEPOR (LM)	0.39 (11)	0.68 (5)
Prism	0.46 (4)	0.68 (6)
COMET-MQM_2021	0.40 (8)	0.67 (7)
BLEU	0.30 (13)	0.65 (8)
YiSi-1	0.42 (7)	0.65 (9)
chrF	0.40 (10)	0.62 (10)
MEE2	0.36 (12)	0.60 (11)
C-SPECpn	0.49 (2)	0.54 (12)
tgt-regEMT	0.5 (1)	0.37 (13)
average	0.42	0.64

Figure A.2: Pairwise accuracy of metrics submitted to the WMT21 Metric Shared Task, reported for the task of ranking system pairs on the TED Chinese→English test-set, using either ref-A (original reference of low quality) or ref-B (extra reference of high quality). For the definition of pairwise accuracy, see Section 3.1. Figure reprint from Freitag et al. (2021).

metric	correlation	metric	correlation
bleurt-20	0.955	COMET-MQM_2021	0.076
COMET-DA_2020	0.949	RoBLEURT	0.075
<u>Prism</u>	0.948	COMET-DA_2021	0.072
bleurt-21-beta	0.947	C-SPEC	0.070
<u>BERTScore</u>	0.947	<u>Prism</u>	0.070
YiSi-1	0.944	C-SPECpn	0.066
RoBLEURT	0.944	COMET-QE-DA_2021-src	0.064
regEMT	0.940	COMET-DA_2020	0.062
COMET-DA_2021	0.939	<u>BERTScore</u>	0.062
sentBLEU	0.936	COMETinho-DA	0.056
<u>chrF</u>	0.924	OpenKiwi-MQM-src	0.051
COMETinho-DA	0.923	YiSi-1	0.049
MTEQA	0.909	COMET-QE-MQM_2021-src	0.047
COMET-MQM_2021	0.902	bleurt-20	0.046
COMET-QE-DA_2021-src	0.898	YiSi-2-src	0.046
COMETinho-MQM	0.880	regEMT	0.043
<u>TER</u>	0.823	bleurt-21-beta	0.039
C-SPEC	0.810	COMETinho-MQM	0.036
OpenKiwi-MQM-src	0.806	chrF	0.021
YiSi-2-src	0.795	regEMT-src	0.009
COMET-QE-MQM_2021-src	0.782	sentBLEU	-0.010
C-SPECpn	0.720	regEMT-baseline	-0.067
regEMT-baseline	0.525	regEMT-baseline-src	-0.067
regEMT-src	0.363	MTEQA	-0.067
regEMT-baseline-src	0.014	<u>TER</u>	-0.125

Figure A.3: Correlations for the Hausa→English translation direction reported on the *newstest2021* dataset. System-level Pearson correlation is reported on the left and the segment-level Kendall-Like correlation on the right. Primary submissions are bolded, and baselines are underlined. Figure reprint from Freitag et al. (2021).

question-answering evaluation task to examine simultaneous (partial) translations. The authors, given a source language question, translate the question word by word into the target language and try to answer it as soon as possible, measuring the quality and timely adequacy of simultaneous translation. Han et al. (2023) propose a framework for handling “explicitation” – an explicit realization of implicit information in the source language that professional translators have thanks to, e.g., cultural knowledge – that explores an automated multilingual QA system to determine whether the explicit realization improves the translation.

B. Auxiliary Results

In this Appendix, we include the numerical results of auxiliary experiments and examples of model outputs.

B.1 COMES

This section is based on the FROM COMET TO COMES — CAN SUMMARY EVALUATION BENEFIT FROM TRANSLATION EVALUATION? (Krubieński and Pecina, 2022) article.

As discussed in Section 3.1.7, due to the cross-validation approach to testing on the SummEval dataset, we trained/fine-tuned several COMES instances. They differ by the exact articles used for training/fine-tuning. Thus, when evaluating on datasets other than SummEval, we evaluate with each instance, reporting mean and standard deviation.

REALSumm results

In Table B.1, we report the system-level Kendall’s τ correlations on the REALSumm corpus (100 articles \times 25 models), annotated by Bhandari et al. (2020). “Score” column is used for metrics that output a single score, the following ones correspond to outputs from each of the COMES heads. From the analysis, we excluded 2 articles that appear in the SummEval dataset.

Metric	LitePyramid SCU				
	Score	Coh	Consistency	Flu	Rel
ROUGE-1 r	0.779				
ROUGE-2 r	0.853				
ROUGE-L r	0.746				
BERTScore r	0.538				
JS-2	0.518				
MoverScore	0.264				
COMET	0.457				
COMES		0.242 \pm 0.05	0.561 \pm 0.07	0.290 \pm 0.02	0.481 \pm 0.05
COMES_MT		0.405 \pm 0.03	0.423 \pm 0.02	0.434 \pm 0.02	0.409 \pm 0.03
COMET_QE	0.745				
COMES_QE		0.264 \pm 0.06	0.592 \pm 0.04	0.309 \pm 0.06	0.490 \pm 0.06
COMES_MT_QE		0.457 \pm 0.05	0.473 \pm 0.04	0.472 \pm 0.04	0.460 \pm 0.05

Table B.1: System-level Kendall’s τ correlations on the REALSumm corpus annotated by Bhandari et al. (2020). The three metrics with the highest correlation in each column are bolded. Table reprint from Krubieński and Pecina (2022).

“Human Feedback” data results

Table B.2 presents the system-level Kendall’s τ correlations on the subset of the test split of the CNN/DailyMail corpus annotated by [Stiennon et al. \(2020\)](#). The columns indicate different evaluation dimensions in the annotated (test) data. In the rows, we include outputs from each of the COMES heads that correspond to evaluation dimensions used in the training data. From the analysis, we excluded 6 articles that appear in the SummEval dataset. In Table B.3, we present the corresponding numbers when evaluating on the subset of the TL;DR corpus annotated by [Stiennon et al. \(2020\)](#) in a similar manner.

Metric		Overall	Accuracy	Coverage	Coherence
ROUGE-1 f		0.647	0.752	0.621	0.464
ROUGE-2 f		0.569	0.699	0.542	0.438
ROUGE-L f		0.595	0.699	0.569	0.412
BERTScore f		0.621	0.725	0.595	0.464
COMET		0.843	0.686	0.817	0.425
COMES	Coh	-0.204 ± 0.05	-0.050 ± 0.04	-0.230 ± 0.05	0.264 ± 0.04
	Con	0.722 ± 0.12	0.630 ± 0.06	0.695 ± 0.12	0.565 ± 0.07
	Flu	0.209 ± 0.10	0.340 ± 0.07	0.186 ± 0.09	0.625 ± 0.07
	Rel	0.774 ± 0.03	0.703 ± 0.04	0.750 ± 0.03	0.627 ± 0.02
COMES_MT	Coh	0.366 ± 0.16	0.403 ± 0.12	0.340 ± 0.16	0.654 ± 0.07
	Con	0.455 ± 0.11	0.418 ± 0.10	0.431 ± 0.12	0.604 ± 0.11
	Flu	0.433 ± 0.12	0.414 ± 0.11	0.407 ± 0.12	0.634 ± 0.06
	Rel	0.379 ± 0.16	0.403 ± 0.12	0.353 ± 0.16	0.654 ± 0.06
COMET_QE		0.922	0.660	0.895	0.477
COMES_QE	Coh	-0.158 ± 0.1	-0.017 ± 0.09	-0.184 ± 0.10	0.305 ± 0.09
	Con	0.714 ± 0.05	0.630 ± 0.05	0.688 ± 0.05	0.544 ± 0.06
	Flu	0.170 ± 0.13	0.272 ± 0.11	0.144 ± 0.13	0.559 ± 0.08
	Rel	0.695 ± 0.07	0.648 ± 0.06	0.669 ± 0.07	0.646 ± 0.04
COMES_MT_QE	Coh	0.480 ± 0.11	0.467 ± 0.09	0.454 ± 0.11	0.668 ± 0.03
	Con	0.528 ± 0.07	0.484 ± 0.08	0.502 ± 0.07	0.638 ± 0.06
	Flu	0.519 ± 0.07	0.480 ± 0.08	0.493 ± 0.07	0.647 ± 0.05
	Rel	0.493 ± 0.09	0.477 ± 0.08	0.467 ± 0.09	0.678 ± 0.02

Table B.2: System-level Kendall’s τ correlations on the subset of CNN/DailyMail corpus annotated by [Stiennon et al. \(2020\)](#). The three metrics with the highest correlation in each column are bolded. Table reprint from [Krubiński and Pecina \(2022\)](#).

Multi_SummEval results

In Table B.4, we report the summary-level (segment-level) Pearson correlations on the subset of Multi_SummEval corpus annotated by [Koto et al. \(2021\)](#). [Koto et al. \(2021\)](#) collected human judgments for *Focus* and *Coverage*, using the Direct Assessment method to collect scores on a continuous scale of 1 to 100. For other metrics, see Table 2 in [Koto et al. \(2021\)](#). For readability reasons, we report only the mean COMES scores and do not report variance.

Metric		Overall	Accuracy	Coverage	Coherence
ROUGE-1 f		0.545	0.000	0.576	0.333
ROUGE-2 f		0.576	0.091	0.606	0.424
ROUGE-L f		0.606	0.061	0.636	0.394
BERTScore f		0.424	-0.121	0.455	0.212
COMET		0.727	-0.061	0.758	0.273
COMES	Coh	-0.058 ± 0.19	0.306 ± 0.15	-0.052 ± 0.18	0.124 ± 0.09
	Con	0.239 ± 0.05	0.082 ± 0.01	0.209 ± 0.05	-0.003 ± 0.05
	Flu	0.227 ± 0.09	-0.106 ± 0.04	0.258 ± 0.09	0.039 ± 0.04
	Rel	0.600 ± 0.12	0.042 ± 0.08	0.630 ± 0.12	0.315 ± 0.08
COMES_MT	Coh	0.682 ± 0.02	-0.100 ± 0.03	0.712 ± 0.02	0.294 ± 0.03
	Con	0.536 ± 0.14	-0.155 ± 0.05	0.567 ± 0.14	0.215 ± 0.09
	Flu	0.561 ± 0.10	-0.161 ± 0.07	0.591 ± 0.10	0.233 ± 0.07
	Rel	0.676 ± 0.03	-0.112 ± 0.03	0.706 ± 0.03	0.282 ± 0.03
COMET_QE		0.545	0.121	0.576	0.394
COMES_QE	Coh	0.088 ± 0.27	0.258 ± 0.14	0.100 ± 0.27	0.173 ± 0.15
	Con	0.206 ± 0.11	0.085 ± 0.06	0.182 ± 0.11	0.012 ± 0.08
	Flu	0.218 ± 0.11	-0.073 ± 0.06	0.248 ± 0.11	0.055 ± 0.06
	Rel	0.533 ± 0.09	0.085 ± 0.07	0.564 ± 0.09	0.315 ± 0.07
COMES_MT_QE	Coh	0.564 ± 0.04	0.048 ± 0.04	0.594 ± 0.04	0.394 ± 0.02
	Con	0.491 ± 0.11	0.012 ± 0.08	0.521 ± 0.11	0.321 ± 0.09
	Flu	0.473 ± 0.11	0.000 ± 0.07	0.503 ± 0.11	0.297 ± 0.10
	Rel	0.555 ± 0.05	0.058 ± 0.04	0.585 ± 0.05	0.385 ± 0.03

Table B.3: System-level Kendall’s τ correlations on the subset of TL;DR corpus annotated by [Stiennon et al. \(2020\)](#). The three metrics with the highest correlation in each column are bolded. Table reprint from [Krubiński and Pecina \(2022\)](#).

Metric	Focus					Coverage					
	de	es	tr	fr	ru	de	es	tr	fr	ru	
COMET	0.82	0.51	0.64	0.47	0.42	0.82	0.54	0.72	0.40	0.45	
COMET_QE	0.29	0.06	0.03	0.01	0.10	0.31	0.09	0.27	-0.03	0.24	
COMES	Coh	0.21	0.03	0.07	0.16	-0.01	0.15	-0.01	-0.05	0.08	-0.07
	Con	0.33	0.11	0.21	0.10	0.14	0.35	0.13	0.30	0.07	0.22
	Flu	0.36	0.05	0.10	0.11	0.08	0.33	0.06	0.10	0.05	0.15
	Rel	0.42	0.15	0.25	0.18	0.12	0.44	0.20	0.38	0.15	0.26
COMES_MT	Coh	0.37	0.13	0.25	0.15	0.08	0.36	0.09	0.31	0.11	0.14
	Con	0.31	0.10	0.20	0.14	0.09	0.30	0.09	0.24	0.09	0.16
	Flu	0.31	0.10	0.21	0.14	0.09	0.30	0.09	0.25	0.09	0.16
	Rel	0.36	0.12	0.25	0.15	0.09	0.35	0.09	0.30	0.10	0.15
COMES_MT_ML	Coh	0.03	-0.01	-0.03	0.13	-0.09	-0.04	-0.04	-0.17	0.10	-0.14
	Con	0.10	0.02	0.01	0.00	0.01	0.10	0.00	0.01	-0.02	0.12
	Flu	0.23	0.02	0.09	0.07	0.01	0.22	0.03	0.08	-0.01	0.01
	Rel	0.36	0.20	0.16	0.15	0.06	0.38	0.25	0.27	0.16	0.23

Table B.4: Summary-level Pearson correlations on the Multi_SummEval corpus annotated by [Koto et al. \(2021\)](#). The three metrics with the highest correlation in each column are bolded. Table reprint from [Krubiński and Pecina \(2022\)](#).

B.2 Examples of Model Outputs

B.2.1 MLASK-MMS

A sample of outputs from the models trained on the MLASK dataset, see Section 5.1. Automatic translations into English are provided for convenience. “System A” is the MMS + SumeCzech + Smooth Labels model, and “System B” is the MMS + SumeCzech + Smooth Labels + Masked Video one.

Reference	<i>Ze století páry rovnou do 21. století. Tak by měla vypadat proměna hlavního brněnského nádraží. Ze současného místa v centru města by se mělo nádraží přesunout o několik set metrů, kde vznikne nový komplex na zelené louce. Jeho podoba by měla být z pera architektů nizozemského ateliéru Benthem Crowwel Architects.</i> From the century of steam straight into the 21st century. This is what the transformation of Brno’s main railway station should look like. From its current location in the city centre, the station should be moved a few hundred metres to create a new green meadow complex. Its design should be from the architects of the Dutch studio Benthem Crowwel Architects.
System A	<i>Nové hlavní nádraží v Brně by mělo přijít na 50 miliard korun. Vítězný ateliér má zkušenosti z Nizozemska, Nizozemska nebo Nizozemska.</i> The new main railway station in Brno should cost CZK 50 billion. The winning studio has experience from the Netherlands, the Netherlands or the Netherlands.
System B	<i>Nové hlavní nádraží v Brně by mělo být hotové do roku 2035. Vítězný ateliér, který má s nádražím bohatou zkušenost, dostal od města peníze za 50 miliard korun.</i> The new main railway station in Brno should be ready by 2035. The winning studio, which has extensive experience with the station, received money from the city worth CZK 50 billion.

Table B.5: Textual summary – MLASK-MMS Example 1. Table reprint from Krubiński and Pecina (2023).

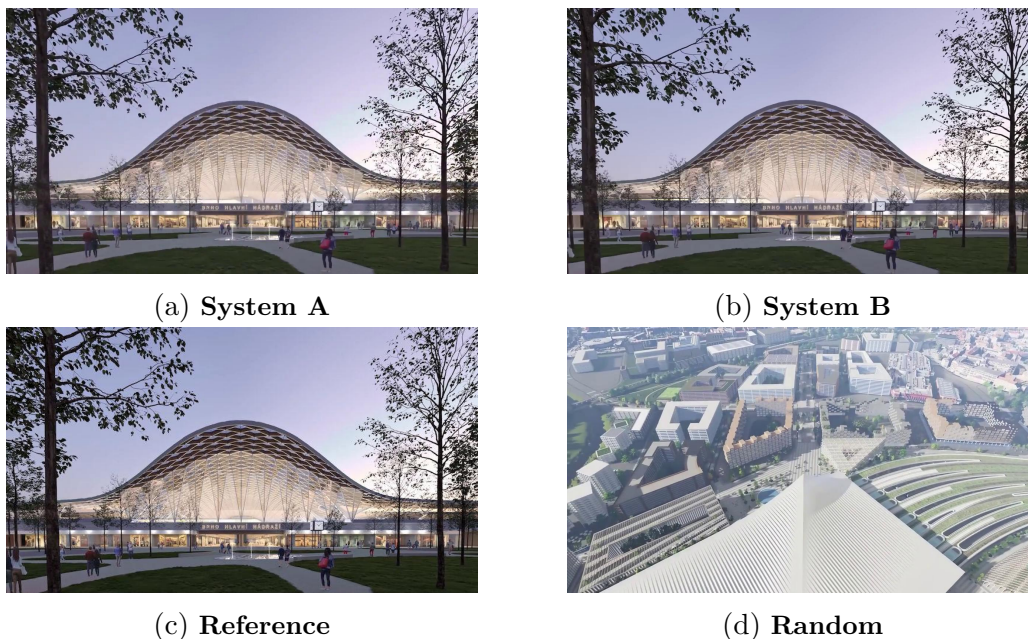


Figure B.1: Pictorial summary – MLASK-MMS Example 1. Figure reprint from Krubiński and Pecina (2023).

Reference	<i>Pětasedmdesátiletý řidič, který při páteční nehodě v Tatcích na Kolínsku usmrtil tříleté dítě a druhé vážně zranil, stejně jako jejich matku, je podezřelý z usmrcení z nedbalosti a z těžkého ublížení na zdraví z nedbalosti. Příčinou nehody byl zřejmě mikrospánek, uvedla v sobotu v tiskové zprávě policejní mluvčí Vlasta Suchánková.</i> The 75-year-old driver, who killed a three-year-old child and seriously injured the other, as well as their mother, in Friday's accident in Tatky, Cologne, is suspected of negligent homicide and grievous bodily harm. The cause of the accident was apparently microsleep, police spokesperson Vlasta Suchánková said in a press release on Saturday.
System A	<i>Dva trestné činy usmrcení z nedbalosti vyšetřují od pátečního odpoledne policisté v Nymburce, kde se srazilo osobní auto s dvěma malými holčičkami. Na dotaz Novinek to ve čtvrtek řekla mluvčí policie Eva Suchánková.</i> Two crimes of culpable homicide have been under investigation since Friday afternoon by police officers in Nymburk, where a car collided with two little girls. Police spokesperson Eva Suchánková told News on Thursday.
System B	<i>Dvě malé holčičky zemřely v pátek odpoledne při nehodě dvou aut na Nymbursku. Jeden člověk byl vážně zraněn, druhý je ve vážném stavu v nemocnici. Právu to řekla mluvčí policie Eva Suchánková.</i> Two little girls died in a two-car accident in Nymburk on Friday afternoon. One person was seriously injured, the other is in a serious condition in hospital. Police spokesperson Eva Suchánková told the law.

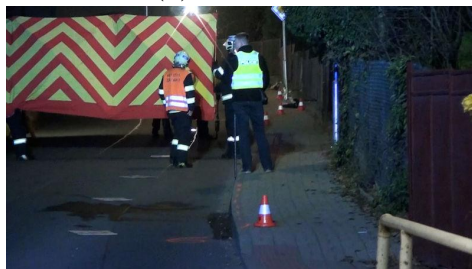
Table B.6: Textual summary – MLASK-MMS Example 2. Table reprint from Krubiński and Pecina (2023).



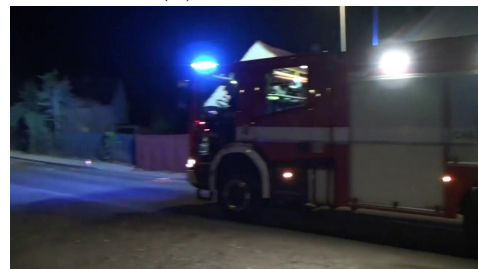
(a) System A



(b) System B



(c) Reference



(d) Random

Figure B.2: Pictorial summary – MLASK-MMS Example 2. Figure reprint from Krubiński and Pecina (2023).

Reference	<p><i>Minulý týden v pátek se v Česku začaly oficiálně prodávat nové chytré hodinky Apple Watch Series 7. Zájem byl tak vysoký, že už po pár hodinách zmizely prakticky všechny hodinky z pultů českých obchodů. A vše nasvědčuje tomu, že si případní zájemci budou muset na další várku ještě dlouho počkat.</i></p> <p>Last week on Friday, the new Apple Watch Series 7 was officially launched in the Czech Republic. Interest was so high that after a few hours practically all watches disappeared from the counters of Czech shops. And all indications are that potential buyers will have to wait a long time for the next batch.</p>
System A	<p><i>V pátek se začaly oficiálně prodávat nové chytré hodinky od společnosti Apple. Zájem o novinku byl tak vysoký, že se v některých případech nepodařilo uspokojit všechny zákazníky, kteří si objednali hodinky s týdenním předstihem.</i></p> <p>On Friday, Apple's new smartwatch officially went on sale. Interest in the new smartwatch was so high that in some cases it was not possible to satisfy all customers who ordered the watch a week in advance.</p>
System B	<p><i>Zájem o novou generaci chytrých hodinek Watch Series 7 byl tak vysoký, že se v některých případech nepodařilo uspokojit všechny zákazníky, kteří si objednali hodinky s týdenním předstihem. Novinka má být daleko lépe než předchůdce – dostala extrémně tenké rámečky okolo displeje.</i></p> <p>Interest in the new generation of Watch Series 7 smartwatches was so high that in some cases it failed to satisfy all customers who ordered a watch a week in advance. The novelty is supposed to be far better than its predecessor – it got extremely thin frames around the display.</p>

Table B.7: Textual summary – MLASK-MMS Example 3. Table reprint from Krubiński and Pecina (2023).



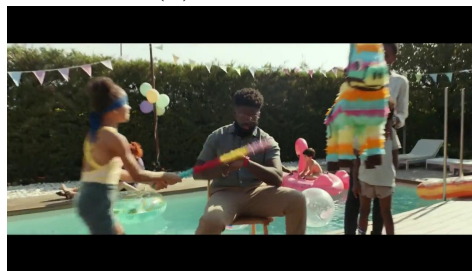
(a) System A



(b) System B



(c) Reference



(d) Random

Figure B.3: Pictorial summary – MLASK-MMS Example 3. Figure reprint from Krubiński and Pecina (2023).

B.2.2 UNMGH

A sample of outputs from the models trained in a multi-task fashion for unified, uni- and multi-modal summarization, see Section 5.2.

Walrus counting from space: How many tusked beasts do you see?



(a) Reference

Thousands of volunteers to count Arctic walrus from space



(b) T5CLIP

Scientists count walrus from space

Satellite image of a Laptev walrus haul-out



(c) BLIP-2

Walrus are heavily dependent on sea-ice, which has been in sharp retreat, leading to increased difficulty for the animals to hunt and rest.



(d) Stable Diffusion 2.1 + Alpaca

Figure B.4: Pictorial summary – M3LS Example 1. Figure reprint from Krubiński and Pecina (2024).

Ireland's Naval Service marks its 75th anniversary



(a) Reference

'I thought the navy was cool and really interesting'



(b) T5CLIP

Covid has ensured that anniversary commemorations will be more subdued than the 50th anniversary celebrations, when foreign navies visited Ireland.

Irish Navy celebrates 75th anniversary



(c) BLIP-2



(d) Stable Diffusion 2.1 + Alpaca

Figure B.5: Pictorial summary – M3LS Example 2. Figure reprint from Krubiński and Pecina (2024).

Man seriously injured his head at waste treatment company, helicopter flew for him



(a) Reference

A worker was injured in a truck at a waste treatment plant in Prague



(b) T5CLIP

A man was injured at a waste treatment company in Prague. He was airlifted to hospital



(c) BLIP-2

A man was injured in a waste treatment company in Prague. He died at the scene



(d) MMS

Man injured at waste treatment plant, airlifted conscious to hospital.



(e) Stable Diffusion 2.1 + Alpaca

Figure B.6: Pictorial summary – MLASK Example 1. Figure reprint from [Kru-
biński and Pecina \(2024\)](#).

I will make the universe accessible to all of you, exulted Branson. Prepare 5 million



(a) Reference

Branson's "a once-in-a-lifetime experience". Take a ride in space with his crew



(b) T5CLIP

Richard Branson became the second 70-yearold to go into space



(c) BLIP-2

The world's richest man has a new era of space travel, Branson and his family are heading to the edge of space



(d) MMS

Virgin Galactic successfully completed its first commercial space flight, marking a major milestone for space tourism.



(e) Stable Diffusion 2.1 + Alpaca

Figure B.7: Pictorial summary – MLASK Example 2. Figure reprint from Kru-
biński and Pecina (2024).

List of Abbreviations

ASR Automatic Speech Recognition. 21–23

ConvNet Convolutional Neural Network. 11, 12, 21, 27, 64, 75

fps frames per second. 55, 84

LLM Large Language Model. 6–8, 23, 38, 39, 45, 75, 77, 78, 80, 90, 143

MSMO Multimodal Summarization with Multimodal Output. 4, 5, 16, 24, 26, 28, 46, 48, 51, 60, 63, 70, 75, 76, 79, 83, 89, 90, 143

MT Machine Translation. 13, 31, 33, 36, 37, 39, 41, 44, 77, 89, 121–125, 143, 146

QA Question Answering. 4, 34, 37, 52, 89, 121, 123, 125, 127

QG Question Generation. 4, 52, 121, 125

RNN Recurrent Neural Network. 7, 8, 13, 21, 26, 27

SOTA State of the art. 21, 36, 37, 75, 90, 125

ViL Vision-and-Language. 4, 5, 10, 75, 77

VMSMO Video-based Multimodal Summarization with Multimodal Output. 26, 27, 48, 63, 84, 89

List of Figures

1.1	Categorization of automatic Text Summarization systems proposed by El-Kassas et al. (2021). Figure reprint from El-Kassas et al. (2021).	5
1.2	An overview of transformer-based cross-modal interactions: a) Early Summation, b) Early Concatenation, c) Hierarchical Attention (multi-stream to one-stream), d) Hierarchical Attention (one-stream to multi-stream), e) Cross-Attention, and f) Cross-Attention to Concatenation. Colors indicate features from separate modalities. Figure reprint from Xu et al. (2023a).	14
2.1	A Multimodal Summarization (MMS) taxonomy proposed by Jangra et al. (2023). The dark-orange nodes coming out of the root (in yellow) represent the segregation based on input, output, and adopted methodology. In contrast, the light-orange nodes following them represent the respective characteristics of the research work on which the works can be distinguished. The teal-colored rectangles in the leaf denote the various categories of each such characteristic. Figure reprint from Jangra et al. (2023).	16
2.2	Example of a multimodal news article from an online publisher (dailymail.co.uk). Three modalities: text, image(s), and video are presented to a user. Each of them brings a new, unique piece of information. While particular modalities may have an inner structure – text can be split into <i>Title</i> , <i>Abstract</i> , and <i>Story</i> , in general, no specific order can be imposed on objects from different modalities.	17
2.3	A multimodal product summarization task proposed by Li et al. (2020a). Figure reprint from Li et al. (2020a).	20
2.4	Examples of visual modality from the MLASK (Krubiński and Pecina, 2023) dataset. Left – the target image. Right – a subset of input video frames, as seen by the model. The target image was modified by removing the watermark in the bottom-right corner.	26
3.1	An overview of evaluation methods by Koto et al. (2022). In the “Manual” column, we see the annotations related to the dimension of human evaluation (Faithfulness, Recall, etc.), a label telling us whether the evaluation was conducted on individual documents (Absolute) or in the context of other texts (Relative), and a taxonomy of methodologies for human evaluation. Figure reprint from Koto et al. (2022)	32
3.2	Illustration of <i>focus</i> and <i>coverage</i> (see Section 3.1.1). Figure reprint from Koto et al. (2022)	33
3.3	Illustration of the Direct Assessment evaluation framework. By switching the position of reference and hypothesis, the same question can be used to collect annotations for <i>focus</i> and <i>coverage</i> . Figure reprint from Koto et al. (2022)	33

3.4	Illustration of the evaluation framework that explores the Likert scale. Figure reprint from Fabbri et al. (2021)	35
3.5	Overview of the QAGS metric. A set of questions is generated based on the summary. The questions are then answered using both the source article and the summary. Corresponding answers are compared using a similarity function and averaged across questions to produce the final QAGS score. Figure reprint from Wang et al. (2020)	38
3.6	An overview of the trainable SummScore metric. Figure reprint from Lin et al. (2022)	40
3.7	Estimator model architecture used in COMES. Source, reference, and hypothesis are all independently encoded with a pre-trained encoder. The pooling layer is used to create sentence embeddings from sequences of token embeddings. In the COMES variant, the last feed-forward layer has 4 outputs corresponding to different summary evaluation dimensions. Dashed lines are used to indicate the reference-less variant. Figure reprint from Krubiński and Pecina (2022).	42
3.8	Screenshot of the annotation tool used to collect human judgments about the quality and usefulness of selected cover frame. For convenience, we translated all text into English. Figure reprint from Krubiński and Pecina (2023).	49
3.9	Values of Cohen’s κ used to measure the inter-annotator agreement on the control batch. Figure reprint from Krubiński and Pecina (2023).	50
4.1	An example (Seznam Zprávy API) of an output from the API call that we used to collect the documents. For clarity, only a subset of retrieved fields is presented.	56
4.2	An annotated screenshot representing one of the articles collected in our experiments. The <image, text> pair in the bottom left corner corresponds to the thumbnail representing the article on the news provider’s main web page, with the “text” field given by the article’s title.	57
4.3	Above – a partial snippet of the bbc.com news hosting website. Below – the reference images (pictorial summaries) collected based on the specific HTML tag (see Section 4.3).	59
4.4	An example of a data instance from the M3LS dataset. For simplicity, all of the retrieved fields are presented, but the actual content is truncated.	61
5.1	An overview of the proposed MMS model for Multimodal Summarization. Figure reprint from Krubiński and Pecina (2023).	64

5.2	Three examples of cosine similarity (y-axis) plots between the numerical features of the reference cover picture and all candidate frames (x-axis) from the video. The examples were chosen manually from the MLASK dataset to present three different video similarity patterns: with a single peak (red), with more than one peak (blue), and with a consecutive sequence of frames having very similar scores (violet). Figure reprint from Krubiński and Pecina (2023).	72
5.3	Overview of the proposed unified approach to MSMO. The visual tokens are appended to the text representation. The generated output includes the textual summary and the <i>index token</i> that indicates which input image (first, second, third, etc.) is picked as the pictorial summary. During training, a mixture of video-based, image-based, and text-only data is used. Figure reprint from Krubiński and Pecina (2024).	76
5.4	Overview of the BLIP-2 model, with an encoder-decoder LLM as the textual component. Figure reprint from Li et al. (2023).	77
5.5	Overview of the BLIP-2 model extended to handle multiple images/frames in the input.	78
5.6	The quality of the visual and textual output during the BLIP-2 fine-tuning, reported on the validation split of the M3LS dataset. We report CosSim and Top-1 Acc metrics for the visual output and ROUGE-L for the textual one.	83
A.1	An illustration of the MTEQA pipeline. One of the MT answers is clearly wrong, one is correct but the other differs with just a single character, raising a question about the choice of the answer-comparison metric. Figure reprint from Krubiński et al. (2021a).	121
A.2	Pairwise accuracy of metrics submitted to the WMT21 Metric Shared Task, reported for the task of ranking system pairs on the TED Chinese→English test-set, using either ref-A (original reference of low quality) or ref-B (extra reference of high quality). For the definition of pairwise accuracy, see Section 3.1. Figure reprint from Freitag et al. (2021).	126
A.3	Correlations for the Hausa→English translation direction reported on the <i>newstest2021</i> dataset. System-level Pearson correlation is reported on the left and the segment-level Kendall-Like correlation on the right. Primary submissions are bolded, and baselines are underlined. Figure reprint from Freitag et al. (2021).	126
B.1	Pictorial summary – MLASK-MMS Example 1. Figure reprint from Krubiński and Pecina (2023).	132
B.2	Pictorial summary – MLASK-MMS Example 2. Figure reprint from Krubiński and Pecina (2023).	133
B.3	Pictorial summary – MLASK-MMS Example 3. Figure reprint from Krubiński and Pecina (2023).	134
B.4	Pictorial summary – M3LS Example 1. Figure reprint from Krubiński and Pecina (2024).	135

B.5	Pictorial summary – M3LS Example 2. Figure reprint from Kru- biński and Pecina (2024).	136
B.6	Pictorial summary – MLASK Example 1. Figure reprint from Kru- biński and Pecina (2024).	137
B.7	Pictorial summary – MLASK Example 2. Figure reprint from Kru- biński and Pecina (2024).	138

List of Tables

3.1	System-level Kendall’s τ correlations with (average) expert annotations for four evaluation dimensions annotated in the SummEval dataset. The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).	43
4.1	Overview of the publicly available datasets explored for Multimodal Summarization. “T” refers to the textual modality, “V” to the video modality, and “I” to the image modality. In the “Language” column, we provide the three-letter ISO 639-2 code. The * symbol indicates a partial match, e.g., only a part of the dataset (usually the test-split) is annotated, or only part of the data (images, but not videos) is released.	54
4.2	Overview of the publicly available datasets explored for Multimodal Summarization, with the corresponding URLs.	54
4.3	Quantitative statistics of the lengths of titles, abstracts, and full texts (measured in the number of tokens) for the MLASK dataset. Q_1 and Q_3 denote the first and the third quartile, respectively. Table reprint from Krubiński and Pecina (2023).	58
4.4	Quantitative statistics of the number of input images (including the target image) in the subset of the English M3LS dataset that we extended with the multimodal target. Table reprint from Krubiński and Pecina (2024).	60
5.1	Evaluation on the dev-set and test-set of MLASK. The figures are averaged over three runs with different seeds. The three highest-scoring systems in each column are bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2023).	69
5.2	System performance on the task of cover picture selection, validated on the subset of the MLASK test-set. Table reprint from Krubiński and Pecina (2023).	72
5.3	Evaluation of the textual output quality on the validation and test splits for each modality-specific dataset. The three highest-scoring systems in each column are bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2024).	81
5.4	Evaluation of the visual output quality on the validation and test splits for each modality-specific dataset. The highest-scoring system in each column is bolded independently for test-set and dev-set. Table reprint from Krubiński and Pecina (2024).	82
5.5	Evaluation of the visual and textual output quality, reported on the post-processed test-split of the MSMO dataset (see Section 4.1).	87

A.1	System-level Pearson correlation for selected metrics used for measuring MT quality with DA human assessment over MT systems using the <i>newstest2020</i> references. Average (<i>avg</i>) is computed over all to-English directions available. A number below the language pair indicates the number of systems considered. Table reprint from Krubiński et al. (2021a).	123
A.2	Examples of the most frequent POS patterns of gold-standard answers in the XQuAD dataset that we explored to create the POS pattern bank. Table reprint from Krubiński et al. (2021a).	123
A.3	System-level Pearson correlation for various variants of the proposed MTEQA metric with DA human assessment over MT systems using the <i>newstest2020</i> references. The average is computed over all to-English directions available. Table reprint from Krubiński et al. (2021a).	124
A.4	Extracted keyphrases and generated corresponding questions for the sentence: “ <i>A cat was caught by spectators at an American football match in Miami Gardens, Florida, after it fell from the stadium’s upper deck.</i> ”, that compare the baseline and keyphrase extraction method based on POS pattern matching and NER. Table reprint from Krubiński et al. (2021a).	125
B.1	System-level Kendall’s τ correlations on the REALSumm corpus annotated by Bhandari et al. (2020). The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).	129
B.2	System-level Kendall’s τ correlations on the subset of CNN/Daily-Mail corpus annotated by Stiennon et al. (2020). The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).	130
B.3	System-level Kendall’s τ correlations on the subset of TL;DR corpus annotated by Stiennon et al. (2020). The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).	131
B.4	Summary-level Pearson correlations on the Multi_SummEval corpus annotated by Koto et al. (2021). The three metrics with the highest correlation in each column are bolded. Table reprint from Krubiński and Pecina (2022).	131
B.5	Textual summary – MLASK-MMS Example 1. Table reprint from Krubiński and Pecina (2023).	132
B.6	Textual summary – MLASK-MMS Example 2. Table reprint from Krubiński and Pecina (2023).	133
B.7	Textual summary – MLASK-MMS Example 3. Table reprint from Krubiński and Pecina (2023).	134

List of Publications

Firstly, we list the publications relevant to this thesis.

MATEUSZ KRUBIŃSKI, PAVEL PECINA: Towards Unified Uni- and Multimodal News Headline Generation. In: *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 437-450, St. Julian's, Malta, 2024. Association for Computational Linguistics. Available at: <https://aclanthology.org/2024.findings-eacl.30.pdf>

MATEUSZ KRUBIŃSKI, PAVEL PECINA: MLASK: Multimodal Summarization of Video-based News Articles. In: *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 910-924, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-eacl.67. Available at: <https://aclanthology.org/2023.findings-eacl.67.pdf>

MATEUSZ KRUBIŃSKI, PAVEL PECINA: From COMET to COMES – Can Summary Evaluation Benefit from Translation Evaluation?. In: *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pp. 21-31, Online, 2022. Association for Computational Linguistics. DOI: 10.18653/v1/2022.eval4nlp-1.3. Available at: <https://aclanthology.org/2022.eval4nlp-1.3.pdf>

MATEUSZ KRUBIŃSKI, ERFAN GHADERY, MARIE-FRANCINE MOENS, PAVEL PECINA: Just Ask! Evaluating Machine Translation by Asking and Answering Questions. In: *Proceedings of the Sixth Conference on Machine Translation*, pp. 495–506, Online, 2021. Association for Computational Linguistics. Available at: <https://aclanthology.org/2021.wmt-1.58.pdf>

MATEUSZ KRUBIŃSKI, ERFAN GHADERY, MARIE-FRANCINE MOENS, PAVEL PECINA: MTEQA at WMT21 Metrics Shared Task. In: *Proceedings of the Sixth Conference on Machine Translation*, pp. 1024–1029, Online, 2021. Association for Computational Linguistics. Available at: <https://aclanthology.org/2021.wmt-1.110.pdf>

Secondly, we list the remaining publications that the author contributed to during his Ph.D. studies.

MATEUSZ KRUBIŃSKI, STEFAN MATCOVICI, DIANA GRIGORE, DANIEL VOINEA, ALIN-IONUT POPA: Watermark Text Pattern Spotting in Document Images. In: *SDU@AAAI-24: The AAAI-24 Workshop on Scientific Document Understanding*, Vancouver, Canada, 2024. Available at: <https://arxiv.org/pdf/2401.05167.pdf>

MATEUSZ KRUBIŃSKI: Basic Arithmetic Properties in the Space of Language Model Prompts. In: *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*, New Orleans, USA, 2023. Available at: <https://mathai2023.github.io/papers/24.pdf>

MATEUSZ KRUBIŃSKI, HASHEM SELLAT, SHADI SALEH, ADAM POSPÍŠIL,
PETR ZEMÁNEK, PAVEL PECINA: Multi-Parallel Corpus of North Levantine
Arabic. In: *Proceedings of ArabicNLP 2023*, pp. 411-417, Singapore, 2023. As-
sociation for Computational Linguistics. DOI: 10.18653/v1/2023.arabicnlp-1.34.
Available at: <https://aclanthology.org/2023.arabicnlp-1.34.pdf>