An End-to-End Multilingual System for Automatic Minuting of Multi-Party Dialogues

Aakash Bhatnagar[&], **Nidhir Bhavsar**[$], **Muskaan Singh**[#] and **Petr Motlicek**[#]
[&] Boston University, Boston, Massachusetts
[$]University of Potsdam, Potsdam, Germany
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
aakash07@bu.edu, bhavsar@uni-potsdam.de,
(msingh,petr.motlicek)@idiap.ch

Abstract

In this paper, we present a pipeline for automatic minuting. This pipeline is an end-to-end system for minuting the multiparty dialogues of meetings. It provides multilingual communication and collaboration, with a specific focus on Natural Language Processing (NLP) technologies: Automatic Speech Recognition (ASR), Machine Translation (MT), Automatic Minuting (AM), Topic Modelling (TM), and Named Entity Recognition (NER). Our summarization model achieves a ROUGE-1 score of 0.45, a BLEU score of 7.069, and a BERT score of 0.673. Our translation model also achieves a high average BERT score of 0.848 across five different languages (de,fr, en, it, and hi). We make our code available at https://github.com/aakash0017/Paclic-summarization-pipeline

1 Introduction

Since the COVID-19 epidemic, a sizeable portion of the working population—particularly those employed in the information technology (IT) sector and academia—has expanded dramatically in virtual meetings. Meetings are, without a doubt, the most important element in fostering teamwork and effective back-and-forth communication. There are numerous Natural Language Processing (NLP) technologies available that give users a complete online interaction experience. The interpretation of these online interactions during remote conferences or meetings is crucial in the globally interconnected world of today.

Summarizing meetings in the form of structured minutes from speech and it can potentially save up to 80% of time.

We realized that generating meeting minutes is a task that is still performed manually and requires a lot of time. Through this paper, we try to solve three major problems encountered during multi-party dialogues via our proposed system.

First, as mentioned above, manually generating minutes consumes much time. Each annotator has to go through hours of recording before writing minutes. Also, each annotator may have a different vocabulary and style, leading to inconsistency in the meeting-minutes format. We try to solve this problem by generating meeting minutes in a consistent format for multi-party conversation(MPC). We incorporate large pre-trained transformer models fine-tuned on MPC meeting datasets.

Second, as globalization is increasing, companies have offices worldwide. Hence, to overcome the language barrier, there is a need for a system that can provide translation of meeting transcripts, meeting minutes, and meeting topics. We provide quick and straightforward translation in five different languages: French(fr), German(de), Russian(ru), Italian(it), and Hindi(hi), allowing businesses to save time and enhance productivity. To further optimize the translation process, we provide isometric translation [1], which generates outputs of length similar to the source length. We believe that isometric translation is the next leap towards a more synchronous auto dubbing process, which can enhance the meeting experience of non-English speaking users.

Lastly, the enormous increase in online meetings and conversations has led to a massive stack of unordered data. It can be cumbersome for users to select the appropriate meeting for their needs. We try to differentiate the generated minutes into multiple segments and align them with corresponding topics derived accordingly. This provides a gist of the meeting without the user listening to the whole recording.

The paper is organized as: Section: 2 brief about the existing work in text and meeting summarization. The proposed methodology is described in Section: 3, where we employ a 3 stage pipeline: ASR; automatic speech recognition, multi-party meeting summarization, isometric translation, and topic segmentation. The experimental setup with the dataset, hyper-parameter settings, and training are in Section:4 with their corresponding results and in Section: 5. Finally, the paper concludes with future prospects in Section: 6.

2 Related Work

In this section, we describe the existing work on text summarization in Section 2.1 and meeting summarization in Section 2.2.

2.1 Text Summarization

Majority of the prior work on meeting summarization investigates how to generate better summaries for news/media article data, such as CNN/Daily Mail [2], Newsroom [3], etc. other tries to summaries scientific documents, such as SciSumm Corpus [4]. However, our paper mainly focuses on meeting summarization which is comparatively a more challenging task. However we do tend to infer some attributes of normal text summarization, which includes both extractive & abstractive methods. Moreover, the topic of meeting summarization, especially automatic minuting has been a demanding research problem across the community [5] [6] [7] and has become a huge part of the text summarization area.

2.2 Meeting Summarization

Since the advent of COVID-19 and the majority of work shifting online, there has been a lot of interest gathering around multi-party dialogue summarization. However, the fundamental idea behind meeting recording summaries has existed for quite some time. [8] suggested a extractive meeting summarization approach using graphs constructed on topical/lexical relations. However, a study conducted by [9] stated the difference between meeting summarization of multi-party transcripts over the Natural Language Generation models for generating unfocused summaries. They proposed multi-modal hierarchical attention across three levels: segment, utterance, and word and suggested a joint model of topic segmentation and summarization. [10]. Next, [10] attempted to pre-train MPC-BERT to find the inherent complicated structure in MPC via crucial interlocutor and utterances. [11] proposes a novel abstractive summary network that adapts to the meeting scenario. It follows a hierarchical structure to accommodate long meeting transcripts and a role vector to depict the difference among speakers.

The aforementioned work, however, creates meeting summaries, whereas our suggested approach attempts to create meeting minutes from the ASR generated transcripts.

3 Proposed Methodology

We propose a pipeline that utilizes a speech-to-text transcription service and a meeting summarization module. Additionally, we provide functionality of topic extraction and isometric translation (German(*de*), French(*fr*), Italian(*it*), Russian(*ru*), and Hindi(*hi*). As depicted in the figure 1, the system accepts a {*.mp3, .mp4*} file consisting of multi-party conversations in English(*en*). Next, we generate ASR output from the input files, which are then utilized by our system to generate meeting minutes. From subsections 3.1 through 3.4, we provide a details overview of various components of our proposed architecture.

3.1 Automatic Speech Recognition (ASR)

For generating optimal transcripts, we use Amazon Transcribe[1], which is a Speech-to-text service offered by Amazon AWS. It holds the largest share in the cloud computing market. Their current English speech recognition model has achieved a word-error-rate (*WER*) of 6.2%. To convert meeting recordings to transcripts, the data must first be uploaded to the Amazon Simple-Storage-Service (*Amazon S3*), which is then used by the Amazon transcribe Speech-to-text API to generate time-sequence order transcripts, with both speaker and utterances stated separately. To handle this, we define a post-processing function that align speaker roles with corresponding utterances, as shown in figure 1. Our system accepts a number of speakers as an argument before applying ASR transcriptions. However, the argument is set to 2 and accepts a maximum value of 10.

3.2 Meeting Summarization

The meeting summarization module generates meeting-minutes from the processed transcripts.

[1] https://aws.amazon.com/transcribe/

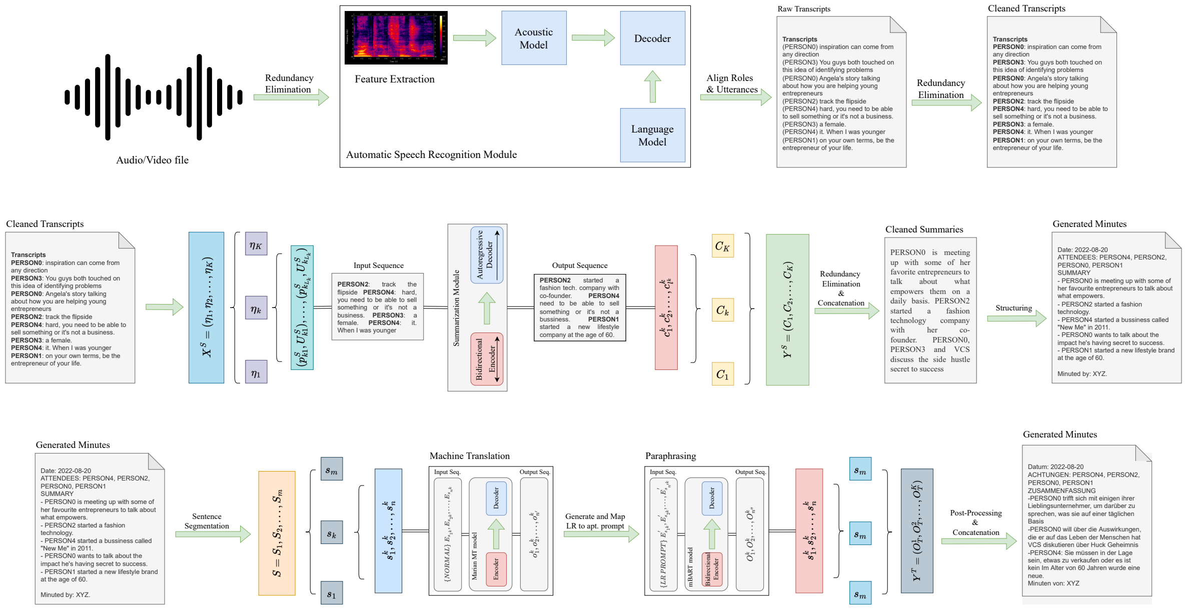
Audio/Video file
Redundancy Elimination
Feature Extraction
Acoustic Model
Decoder
Language Model
Automatic Speech Recognition Module
Align Roles & Utterances
Raw Transcripts
Transcripts
(PERSON0) inspiration can come from any direction
(PERSON3) You guys both touched on this idea of identifying problems
(PERSON0) Angela's story talking about how you are helping young entrepreneurs
(PERSON2) track the flipside
(PERSON4) hard, you need to be able to sell something or it's not a business.
(PERSON3) a female.
(PERSON4) it. When I was younger
(PERSON1) on your own terms, be the entrepreneur of your life.
Redundancy Elimination
Cleaned Transcripts
Transcripts
PERSON0: inspiration can come from any direction
PERSON3: You guys both touched on this idea of identifying problems
PERSON0: Angela's story talking about how you are helping young entrepreneurs
PERSON2: track the flipside
PERSON4: hard, you need to be able to sell something or it's not a business.
PERSON3: a female.
PERSON4: it. When I was younger
PERSON1: on your own terms, be the entrepreneur of your life.
Cleaned Transcripts
$X^S = (\eta_1, \eta_2, \ldots, \eta_K)$
η_K
η_k
η_1
$(p^S_{k1}, U^S_{k1}), \ldots, (p^S_{k_{L_k}}, U^S_{k_{L_k}})$
Input Sequence
PERSON2: track the flipside PERSON4: hard, you need to be able to sell something or it's not a business. PERSON3: a female. PERSON4: it. When I was younger
Summarization Module
Autoregressive Decoder
Bidirectional Encoder
Output Sequence
PERSON2 started a fashion tech. company with co-founder. PERSON4 need to be able to sell something or it's not a bussiness. PERSON1 started a new lifestyle company at the age of 60.
$c^k_1, c^k_2, \ldots, c^k_{l^k}$
C_K
C_k
C_1
$Y^S = (C_1, C_2, \ldots, C_K)$
Redundancy Elimination & Concatenation
Cleaned Summaries
PERSON0 is meeting up with some of her favorite entrepreneurs to talk about what empowers them on a daily basis. PERSON2 started a fashion technology company with her co-founder. PERSON0, PERSON3 and VCS discuss the side hustle secret to success
Structuring
Generated Minutes
Date: 2022-08-20
ATTENDEES: PERSON4, PERSON2, PERSON0, PERSON1
SUMMARY
- PERSON0 is meeting up with some of her favourite entrepreneurs to talk about what empowers.
- PERSON2 started a fashion technology.
- PERSON4 started a bussiness called "New Me" in 2011.
- PERSON0 wants to talk about the impact he's having secret to success.
- PERSON1 started a new lifestyle brand at the age of 60.
Minuted by: XYZ.
Generated Minutes
Sentence Segmentation
$S = S_1, S_2, \ldots, S_m$
s_m
s_k
s_1
$s^k_1, s^k_2, \ldots, s^k_n$
Machine Translation
Input Seq.
Output Seq.
$\{NORMAL\}\ E_{s_1k}, E_{s_2k}, \ldots, E_{s_nk}$
Marian MT model
Decoder
Encoder
$o^k_1, o^k_2, \ldots, o^k_{n'}$
Generate and Map LR to apt. prompt
Paraphrasing
Input Seq.
Output Seq.
$\{LR\ PROMPT\}\ E'_{s_1k}, E'_{s_2k}, \ldots, E'_{s_nk}$
mBART model
Decoder
Bidirectional Encoder
$O^k_1, O^k_2, \ldots, O^k_{n''}$
$s^k_1, s^k_2, \ldots, s^k_n$
s_m
s_m
s_m
$Y^T = (O^1_T, O^2_T, \ldots, O^K_T)$
Post-Processing & Concatenation
Generated Minutes
Datum: 2022-08-20
ACHTUNGEN: PERSON4, PERSON2, PERSON0, PERSON1
ZUSAMMENFASSUNG
-PERSON0 trifft sich mit einigen ihrer Lieblingsunternehmer, um darüber zu sprechen, was sie auf einer täglichen Basis
-PERSON0 will über die Auswirkungen, die er auf das Leben der Menschen hat VCS diskutieren über Huck Geheimnis
-PERSON4: Sie müssen in der Lage sein, etwas zu verkaufen oder es ist kein Im Alter von 60 Jahren wurde eine neue.
Minuten von: XYZ

Figure 1: Displays the entire architecture of our proposed systems.

The task of automatic-minuting differs distinctively from the summarization task. Minuting is primarily concerned with capturing and providing a third-person perspective of important points raised throughout the meeting, whereas summarizing is more concerned with delivering a piece of concise information and not reflecting small details. Our pipeline overcomes one major drawback of manual minuting; that the minutes format and language vary across different annotators.

The meeting summarization module is divided into three main parts. First, we start by preprocessing input transcripts, apply redundancy elimination and segmentation. Next, we apply our meeting summarization model. Finally, we filter the output using an unsupervised redundancy elimination method to obtain the processed minutes.

The majority of dialogue summarization system lacks the ability to refrain from redundancies. Besides that they are also limited to a specific length of input sequences for accurate text generation. Our proposed system tries to tackle these issues using the Redundancy Elimination and Segmentation module. We employ some handcrafted rules and pre-processing techniques to process the input utterances obtained from previously generated transcripts. First, text cleaning procedures are used to get rid of any repeats, pauses, and interruptions in the text. These utterances are then filtered using a custom stopwords we define from publicly available meeting summarization corpuses like AMI [12] & ISCI [11]. Next, we utilize some brute force approach to slice the non-redundant transcripts to address the limitation of length constraints of input sequences. Currently our system support segmentation for varying token lengths of *512*, *768* and *1024* respectively.

We use a finetuned BART-large model [13] [2] for our primary summarization task. BART is a denoising autoencoder for pretraining sequence-to-sequence models. The model is trained by using arbitrary denoising functions to distort text and then instructing it to recreate the original content. Using BART provides the ability to use bi-directional attributes when operating on sequence generation tasks which makes it useful for abstractive text summarization. While BERT cannot adopt a bidirectional mechanism for sequence generation, BART exploits the GPT-2 architecture for predicting the following words with the help of words encountered previously in the current sequence. Hence, we primarily test the pipeline with various BART-based setups. However, we majorly experiment with a fine-tuned version of BART trained simultaneously on XSum [14] & SAMSum[15] datasets.

The generated summaries contain a sufficient amount of information, although they are not entirely adequate. There might be an inclusion of casual discussion or other unnecessary information. This problem is addressed with TextRank. Based on our experimentations, we found out that from

[2] https://huggingface.co/lidiya/bart-large-xsum-samsum

the whole report, the model typically catches 15% of trivial and unnecessary information. We rank the summary lines in increasing order of their importance and exclude out bottom 15% of the lines to obtain a "gold span" of the summary. To further compress the summaries, we add appropriate pronouns, eliminate grammatical inconsistencies wherever possible, and filter the final chain of conversation threads by excluding unnecessary words using stopwords set that we internally develop by observing the generated summaries.

3.3 Isometric Machine Translation

The Machine translation module provides set up the capability to generate transcripts, minutes, and topics in five different languages *de, fr, it, ru, hi*. For all these languages, we provide a user with isometric translation output. Isometric MT is the concept of generating translation that falls within the source length range of $\pm 10\%$. This feature helps to generate synchronous outputs upon text-to-speech conversion. For implementing isometric translation, we develop a multitask learning model similar to [16]. We use fine-tuned OPUS-MT [17] model for translation and fine-tuned mBART [18] for paraphrasing. However, our isometric translation module works best for French, German, and Russian languages as we implement a paraphrasing model to enhance the vocabulary. Hindi and Italian translation does not contain a paraphrasing model, but the use of prompt engineering techniques enables them to achieve a high BLEU score and BERT score.

3.4 Topics Modelling

DeepCon also provides a feature for automatic topic extraction based on Named Entity Recognition that extracts the top-k repeating n-grams from the transcripts. We use Yake[3] library for extracting named entities. Our system can also translate the keywords into the five different languages *de, fr, it, ru, hi*. We intend to generate these topics or keywords in order to provide a comprehensive abstract view of meeting discussions with the generated minutes.

4 Experimental Setup

In this section, we describe dataset details in section 4.1, hyper-parameter setting in section 4.2 and training procedures in section 4.3

4.1 Dataset

As stated, our summarization module utilizes a BART model fine-tuned on both XSum and SAMSum datasets. XSum dataset includes short summaries of articles and discussions, whereas SAMSum is a multi-party meeting conversation dataset usually comprising casual and friendly conversations. Training model on these two datasets allows it to grasp summarization both at the syntactic and morphological levels. For evaluating our proposed summarization models, we used the publicly available ELITR Minuting Corpus [19]. The corpus is divided into 3 subtasks. However, we used the Task-A dataset with the dataset distribution of $85, 10$, and 25 instances for train, validation, and test set, respectively. Each instance comprises i) a meeting transcript and ii) one or more than one human annotated meeting minutes. Table 1 shows the statistics of the dataset that we have used in experimentation.

Next, for the isometric translation module, we experimented with multiple datasets across previously specified languages for both the machine translation and paraphrasing modules. For machine translation, we majorly use the Multilingual Speech Translation Corpus (MuST-C) [20]. We also utilise the Statistical Machine Translation Dataset (WMT) [21] for German (de) & Russian (ru) text inputs. Additionally for translating Hindi (hi) we use the IIT-B Hindi-English Corpus [22]. Next, we use a combination of Opusparcus [23] and PAWS-X [24] datasets for most of our Paraphrase training tasks, However, due to unavailability of PAWS-X dataset for Russian (ru), we utilize the Tapaco [25] dataset which is a sub-extracted paraphrase corpus derived from the Tatoeba database [26]

4.2 Hyper-parameter Settings

We used 4 Tesla V100-PCIE GPUs for all experiments with a memory size of 32510 MiB each. Due to resource constraints, we train each of our models both for summarization and isometric machine translation for 1 epoch, each with a batch size of 32. Our fine-tuned models are trained on a learning rate configuration of 2×10^{-5}. For finetuning the underlying summarization model, we use the following configurations: *'max input length' = 512*, *min target length = 128*. Next, for the isometric machine translation module for both our machine translation and paraphrasing model training, we implement the *AdaFactor* optimizer, which inter-

[3] https://github.com/LIAAD/yake

Table 1: Represent the various statistics calculated on both the SAMSum and ELITR datasets. This includes the No. of dialogues, No. of turns, No. of speakers, No. of average turn lengths, Length of dialogues, Summary lengths, and Percentage of compression.

Datasets	# diag.	# turns	# speakers	avg. turn len.	# len. of diag.	# summary len.	% comp.
SAMSum	16.4K	11.2	2.4	9.1	124	23.4	82.12
ELITR	124	254.4	5.8	9.7	8890.8	387	95.65

nally adjusts the learning rate based on the scale parameter and relative/warmup steps.

4.3 Training

In this section, we discuss all experiments performed for our proposed system. We experimented with various automatic speech recognition (ASR) models for generating MPC transcripts. This includes Wav2Vec[4] [27] model trained on the MInDS-14 [28] dataset. We used the Word-Error-Rate (WER) to evaluate these models. However, most of our trained models generated the speech-to-text output with reasonably high WER scores, runtime, and Samples per Second during testing. Additionally, the transcripts that were generated appear cluttered with extra incomplete content. Thus finally, we decided to use the Amazon Transcribe service to generate meeting transcriptions for further processing.

Next, we experiment with multiple summarization models using T5 [29], Pegasus [30], RoBERTa2RoBERTa [31], distilBART [32], etc. However, the BART-based pipeline performed better than the rest. Table 2 represent the scores evaluated on the ELITR Task-A test dataset. Our experiments include fine-tuning these pre-trained models on various summarization datasets. This includes, CNN/DailyMail, XSUM, SAMSUM and AMI Corpus.

We also implement a singleton MT5 model that performs translation and paraphrasing of 5 supported languages using the prompt engineering method. In this approach, we use two additional prompts combined with length prompts: 1) Translation and 2) Paraphrasing. The translation prompt signifies that the model will translate the given input, and the paraphrasing prompt signifies that the model will generate isometric sentences from the translated sentences. We use the MUST-C dataset for translation and PAWS-X and Topaco for paraphrasing. However, the sentences generated by this MT5 model are very redundant and non-contextual.

Table 2: Performance of different baseline models considered during experimentation. This includes Rouge-1, Rouge-WE, BLEU score, BERT-F1 score, TF-IDF score.

Models	R1	RWE	BLEU	BERT	TFIDF
BART	0.297	0.162	2.907	0.563	0.19
DistilBART	0.375	0.205	6.535	0.620	0.25
T5	0.406	0.229	6.278	0.615	0.31
Ours	**0.45**	**0.298**	**7.068**	**0.673**	**0.38**

Next, we adopt a prompt-based few-shot learning strategy for the paraphrasing task. The model utilizes a small sample of the training dataset(approx 500) and then tries to integrate the derived model with the predictions obtained from the MT model. The comparative scores achieved by assessing using the same technique are listed in table **??**. The few-shot model can adequately constrain the output length while preserving the MT's semantical aspects.

5 Results and Analysis

As said earlier, our system accepts total speakers in the range $\{2, 10\}$. Also, providing an exact count of total speaker value shows that it helps the Amazon Transcribe model to generate the best results and align each speaker utterance with its audio counterpart.

We evaluate our proposed summarization model based on the following metrics. This includes i) ROUGE-N; to match n-grams between system predictions and target gold spans. ii) ROUGE-WE; Since ROUGE-N is extremely biased toward lexical similarities, we attempt to compare the projected summaries using the word embeddings based ROUGE as described in [33]. iii) BLEU; though preferred for evaluating machine translation output, we use this metric to calculate the quality of generated summaries. iv) BERT-score; it calculates the semantic relatedness by aligning the sentence representation of both reference and hypothesis using the cosine similarity. v) TF-IDF; works by calculating the importance of each word/token in generated output based on its occurrence in the doc-

[4] https://huggingface.co/facebook/wav2vec2-base-960h

language	model	dataset	BLEU Score	BERT Score	Length Ratio	Length Range
de	OPUS-MT	MuST-C + WMT	**42.3**	**0.85**	**1.087**	49.81
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	29.1	0.83	1.04	50.55
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	29.9	0.83	1.05	**51.95**
it	OPUS-MT	MuST-C	**34**	**0.84**	**1.045**	**57.032**
fr	OPUS-MT	MuST-C	**44.8**	**0.87**	1.08	49.6
	OPUS-MT+ MT5	MuST-C	42.3	0.85	1.12	51.3
	OPUS-MT + MT5	MuST-C	38	0.86	1.11	46.4
	OPUS-MT + few short mBART	MuST-C + Opusparcus + Paws-X	40.9	0.85	**1.03**	57.33
	OPUS-MT + mBART	MuST-C + Opusparcus + Paws-X	41.2	0.85	1.04	**61.81**
ru	OPUS-MT	MuST-C + WMT	**22.7**	**0.84**	**1.005**	54.517
	OPUS-MT + few short mBART	MuST-C + WMT + Opusparcus + Paws-X	20.8	0.82	0.95	58.934
	OPUS-MT + mBART	MuST-C + WMT + Opusparcus + Paws-X	21.7	0.83	0.967	**62.475**
	MT5	MuST-C + WMT	5.6	0.76	0.732	19.3
hi	OPUS-MT	IITB-En-hi	**11.9**	**0.84**	**0.941**	**42.521**

Table 3: Evaluation scores of various experiments. In this table, we state the language-wise experiments along with the datasets used

ument. Table 2 shows performance analysis of our proposed summarization models and its comparison to various other summarization models. As is evident, our suggested approach produces better results when compared to the other models, and by a significant margin. This indicates that our generated meeting minutes are more accurate regarding Grammatical Correctness and Fluency than the other recent summarization models.

We use BLEU, BERT-score and length compliance metrics to evaluate isometric translation outputs. As mentioned earlier, a statistical method evaluates based on n-grams in translated and reference text and rates the quality of the predictions. BERT-score and Length Compliance metrics are specially designed for the task of isometric MT. The Length Compliance metrics comprise 2 measures, a) length ratio, calculated by matching the length of predicted text against the gold-span targets, and b) length range, which measures the percentage of sentences that falls within the ±10 ideal span of the length-ratio. Table 3 represents the evaluation scores obtained by training various models during experimentations across the previously mentioned MuST-C and IIT-B test datasets. First, the best-performing models for the isometric task for each source language have a lower BLEU score. This suggests that the isometric constraints can affect the calculated BLEU score since it is character-dependent. However, emulating the predicted MT text via the paraphrase model suggests a higher BLEU score. This is because the paraphrasing module modulates the sentence length to conform to the interchangeable vocabulary.

6 Conclusion

The proposed pipeline efficiently handles audio/video files and generates meeting minutes, translations and topics. However, the pipeline does not extract any feature from the video. We believe that using video frames and fusing them with the embeddings of ASR output can generate some quality results. By Introducing the multi-modality aspect, we can further leverage the essential information the video provides. The task of multi-modal fusion poses a significant challenge, and thus we hope to counter it in our upcoming projects. This pipeline can also be extended as an API service for developers to incorporate Auto-minuting functionality in their systems.

7 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

[1] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *ACL*, 2022.

[2] Ramesh Nallapati, Bing Xiang, and Bowen Zhou. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023, 2016.

[3] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *CoRR*, abs/1804.11283, 2018.

[4] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*, 2019.

[5] Lu Wang and Claire Cardie. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 45–53, Philadelphia, Pennsylvania, U.S.A., June 2014. Association for Computational Linguistics.

[7] Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Yun-Nung Chen and Florian Metze. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 377–381, Montréal, Canada, June 2012. Association for Computational Linguistics.

[9] Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. *CoRR*, abs/2106.01541, 2021.

[11] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. End-to-end abstractive summarization for meetings. *CoRR*, abs/2004.02016, 2020.

[12] Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *CoRR*, abs/1809.05715, 2018.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[14] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *CoRR*, abs/1808.08745, 2018.

[15] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237, 2019.

[16] Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. Hierarchical multi-task learning framework for isometric-speech language translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 379–385, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[17] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.

[18] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020.

[19] Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondrej Bojar. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. 2022.

[20] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[22] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT bombay english-hindi parallel corpus. *CoRR*, abs/1710.02855, 2017.

[23] Mathias Creutz. Open subtitles paraphrase corpus for six languages. *CoRR*, abs/1809.06142, 2018.

[24] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. *CoRR*, abs/1908.11828, 2019.

[25] Yves Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association.

[26] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, 2018.

[27] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020.

[28] Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal Lis, Eshan Singhal, Nikola Mrksic, Tsung-Hsien Wen, and Ivan Vulic. Multilingual and cross-lingual intent detection from spoken data. *CoRR*, abs/2104.08524, 2021.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[30] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777, 2019.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[32] Sam Shleifer and Alexander M. Rush. Pre-trained summarization distillation. *CoRR*, abs/2010.13002, 2020.

[33] Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal, September 2015. Association for Computational Linguistics.