

UrduMASD: A Multimodal Abstractive Summarization Dataset for Urdu

Ali Faheem, Faizad Ullah, Muhammad Sohaib Ayub, Asim Karim
Lahore University of Management Sciences (LUMS)



LREC-COLING  2024

Introduction

- Multimodal summarization integrates diverse content types—text, audio, and visual elements—to produce concise and informative summaries.
- This approach is essential in an era where digital platforms like YouTube and Instagram host vast amounts of multimodal content, necessitating effective summarization tools to enhance user experience.
- Such summarization techniques not only improve content accessibility and discoverability but also boost engagement by providing users with succinct, relevant content overviews.

Related Work and Dataset Comparison

- Overview of multimodal summarization methodologies.
- Comparison of resources available for high-resource versus low-resource languages.
- Technological advancements and their gaps in Urdu summarization.

| Dataset | Total Docs | Input Modality | Language | Avg Doc Length | Avg Sum Length |
|-----------|------------|------------------|----------|----------------|----------------|
| CNN | 92,539 | Text | En | 760.50 | 45.70 |
| DailyMail | 219,506 | Text | En | 653.33 | 54.65 |
| XSum | 226,711 | Text | En | 431.07 | 23.26 |
| How2 | 79,114 | Text+audio+video | En | 282.57 | 32.99 |
| UrduMASD | 15,374 | Text+audio+video | Ur | 363.14 | 50.92 |

Table: Comparison of UrduMASD with other benchmark abstractive summarization datasets

UrduMASD Dataset Overview



Title

پشتون موسیقی، رباب کے ماتے پڑنے ستر

Pashtun Music, The declining sounds of Rubab

Transcription

پشتون موسیقی کا شن شاہ کیلئے والے عال مسیقی رباب گٹن سے ملتا جلتا لی موسیقی ہے لیکن دونوں کی آواز اور شکل میں واضح فرق پایا جاتا ہے۔ رباب کی تین قسمیں ہوتی ہیں بڑے سائس کارب باقیس چھوٹے اور بڑے تاروں پر مشتمل ہوتا ہے جبکہ دو تاروں میں رباب کے ایک ماہر اپنے تھے اس کا رباب بچتا مجھے بہت اچھا لگا اور پھر مجھے بھی سیکھنے کا شوق پیدا ہوا میں ذب جانی والی رباب کی بہ عامی محفلوں میں سائس کی مقبولیت میں اہم کردار ادا کرتی رہی ہیں لیکن گزرتی آگے عرصہ سے یہ پروگرامات نہ ہونے کے برابر ہے اور شاید اسی وجہ سے رزگاری حیثیت حاصل ہے ہاں تاریخدانوں کا کہنا ہے کہ جنوبی ایشیا میں رباب کی ابتدا سب سے پہلی ہزار اور وہ بھی رقیق تاروں سے تھی جو ایشیا اور پاکستان کا مشہور جن سے اس سار کو استعمال کیا مختلف تاروں کے سائس میں سے اس کو ہمارے پورے زمین میں اس ساروں کا ایک فرق کا پائیدار ہے۔ پشتون علاقوں میں بڑھتے ہوئے رحمان اور شرد پسندی کے باعث پشتون علاقوں میں روایتوں کی اپنی سلفت سے لا علمی کی وجہ سے اسے اپنی ساروں کی مقبولیت محوم ہوتی موسیقی کے مطالعہ رباب کر سہ میں اتنا ایک سائنٹسٹک عمل ہے اس میں جو چھوٹے تار ہوتے ہیں اس کو بڑے تاروں سے جوڑنا پڑتا ہے اور شاید یہی وجہ ہے کہ یہ ایک مشکل فن بھی سمجھا جاتا ہے۔

Known as the Shin Shah of Pashto music, he is an expert musician in both the rabab and guitar. However, there is a clear distinction in their sound and appearance. The rabab comes in three types: the large-sized Saisorabab, the 21-string Carbab, and the smaller one with both big and small strings. Inspired by the mastery of a rabab player, I developed a keen interest in learning to play the rabab. These public gatherings featuring the Dz Rabab have played a significant role in enhancing its popularity. However, for some time now, such programs have dwindled, possibly contributing to its diminishing presence. Historians claim that the origin of the rabab in South Asia is attributed to Rafeeq Ghazni, who used this instrument in India and Pakistan. Among various string instruments, it has gained immense recognition not just in our region, but has seen an increase in its popularity worldwide, primarily due to its unique sound and charm. In the Pashtun regions, it has a deep connection to their cultural heritage, and people admire it because of its unexplored musical qualities. Learning to play the rabab is an intricate process; it involves connecting small strings to larger ones, which may be why it's considered a challenging art.

Summary (Description)

پاپ میوزک کے بڑھتے ہوئے رحمان اور شدت پسندی کے باعث پشتون علاقوں میں رباب جیسے روایتی آلات موسیقی کا استعمال کم ہو رہا ہے۔

Due to the increasing popularity of pop music and a preference for intensity, the use of traditional musical instruments like the rabab is diminishing in Pashtun regions.

- Components of the dataset: Videos, audio files, transcripts, summaries.
- Data diversity and source selection criteria.

Figure: An illustrative example of UrduMASD.

Data Collection and Preprocessing

- Sources: Urdu news channels on YouTube.
- Use of ASR for transcription and preprocessing challenges.
 - Whisper Medium Urdu
 - Mixed Language Transcripts. Used *langdetect* library to filter other languages
- Filter on Video length

Dataset Statistics

- Total collections: 15,374 videos, audio, titles, transcripts, and summaries.
- Average transcript length: 363.14 words; average summary length: 50.92 words.
- Data diversity: Covers a broad range of topics and genres.
- Comparison with How2 (text+audio+video, English): How2 has an average doc length of 282.57 words and average summary length of 32.99 words.

Intrinsic Evaluation of Dataset

- Metrics used for intrinsic evaluation: Abtractivity (ABS), Compression (CMP), Redundancy (RED), and Semantic Coherence (SC).
- UrduMASD exhibits high ABS (0.958) indicating a high level of abstraction in summaries.
- The dataset shows low RED (0.111), suggesting minimal repetition and a higher unique content ratio.
- SC at 0.527 reflects challenges in semantic coherence, possibly due to the model's comprehension of Urdu.

| Dataset | ABS | CMP | RED | SC |
|----------------|------------|------------|------------|-----------|
| How2 | 0.479 | 0.858 | 0.187 | 0.949 |
| UrduMASD | 0.958 | 0.732 | 0.111 | 0.527 |

Table: Intrinsic Evaluation of UrduMASD and How2 Datasets

Models Used for Urdu Summarization

- **mT5 (Multilingual T5):** Leveraged for its capacity in transfer learning, mT5 is a variant of the T5 model pre-trained on a diverse set of languages, including Urdu. It is used for text-based summarization.
- **MLASK: Multimodal Article Summarization Kit** The model architecture comprises three main components: a Feature Encoder incorporating text, video, and frame encoders; a Cross-modal Interaction Module that blends visual and textual representations; and a decoder.

Experimental Setup

- Utilized mT5 and MLASK models for Urdu video-based summarization, leveraging their transfer learning capabilities for low-resource languages.
- Employed three input configurations to evaluate the impact of multimodal data:
 - Text-only transcriptions as baseline input.
 - Enhanced input with visual data integration using image captions.
 - MLASK model combining text and video features with attention mechanisms.

Results and Discussion

- ROUGE scores indicate the quality of summaries with and without multimodal data:
 - Text-only mT5 models show lower scores compared to multimodal MLASK.
 - Incorporation of visual information led to a 2.6% improvement in ROUGE scores.
- Multimodal integration demonstrates enhanced summary quality, capturing more nuanced information from the data.

| Modality | Model | Input | ROUGE-1 | ROUGE-2 | ROUGE-L |
|------------|-----------|------------------------------|---------|---------|---------|
| Text | mT5-small | Transcripts | 23.91 | 5.59 | 17.01 |
| Text | mT5-base | Transcripts | 22.12 | 4.80 | 15.96 |
| Text | mT5-small | Transcripts + Image Captions | 26.89 | 8.00 | 19.69 |
| Text | mT5-base | Transcripts + Image Captions | 23.27 | 6.10 | 17.03 |
| Multimodal | MLASK | Transcripts+Video | 23.86 | 6.34 | 17.40 |

Table: ROUGE-1, ROUGE-2, and ROUGE-L evaluation scores of mT5-small, mT5-base, and MLASK on different experimental settings

Conclusion

- Introduced UrduMASD, the first Urdu video-based multimodal dataset, which enriches the scope of NLP research for low-resource languages.
- Demonstrated the effectiveness of multimodal inputs, with visual data improving summary quality as evidenced by a 2.6% increase in ROUGE scores.
- Highlighted the necessity for specialized NLP models to better understand and process Urdu, suggesting a future direction towards dedicated model pretraining.
- Emphasized the potential of UrduMASD to pave the way for advancements in automatic summarization, particularly in the multimedia domain.

Thank You

Thank you for your attention!