Nagaoka University of Technology at TRECVID 2022: Video to Text

Team kslab at Nagaoka University of Technology
Mutsuki Ishii
Toshichika Mashimo
Takashi Yukawa

Outline

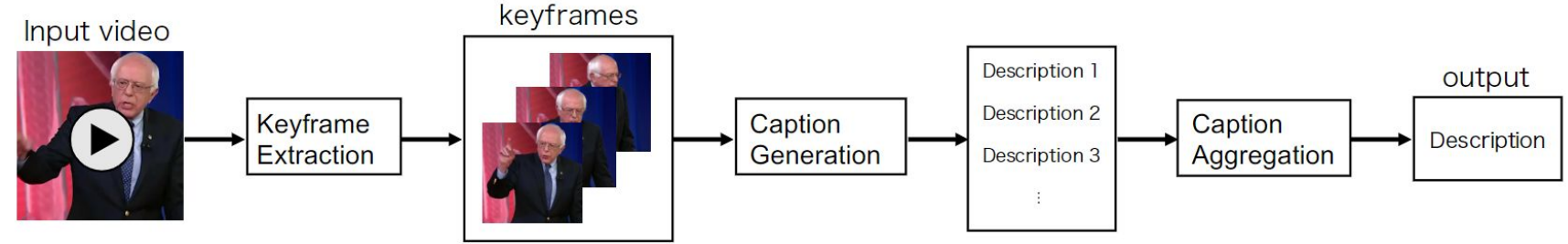
1. Introduction
2. Previous Work
3. Issues & The Target of This Year
4. Captioning Phase
5. Aggregation Phase
6. Result
7. Observatoins
8. Conclution

Introduction

- The method, which uses the whole video frame, is highly accurate but has some problems:
  - A lot of computational resources to construct the system.
  - A long time to generate sentences.
- Our final goal:
  - To reduce the computational resources required to build the system.
  - To generate sentences in a short time.
  - To generate captions with high accuracy.

Previous Work

1. We proposed an architecture consisting of three phase:
    1. Keyframe extraction
    2. Caption generation
    3. Caption aggregation



Previous Work

- Select 7 frames from the video using KTS (Kernel Temporal Segmentation)[1].
- Generate explanatory caption from the extracted keyframes from the NIC model[2].
- Summarize captions into a single sentence by aggregating them using Lexrank[3].

[1] D. Potapov, M. Douze, Z. Harchaoui, andC. Schmid. Category-specific video summariza-tion. InEuropean Conference on Computer Vision(ECCV), volume 8694 ofLecture Notes in ComputerScience, pages 540–555. Springer, Sep 2014.
[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan.Show and tell: A neural image caption generator.Computing Research Repository, arXiv:1411.4555,2015.
[3] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, Vol. 22, No. 1, pp. 457−479, 2004.

Issue & The Target of This Year

- The previous system had the problems.
  - A lack of representation regarding the subject of the video depicted in the caption.
  - The inclusion of words that are irrelevant to the content of the video.

- This year’s targets
  - To generate more expressive descriptions in the captioning phase.
  - Appropriate word selection in the aggregation phase.

Captioning Phase

- Captioning Phase to generate captions in OFA[4].

- OFA is a framework that uses an encoder and decoder based on the Seq2Seq model combined with the transformer.

[4] P. Wang et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. Computing Research Repository, arXiv2202.03052, 2022.

Aggregation Phase

- In the word-by-word method, a extracting phrases and form captions from the aggregated results using Lexrank for each phrase.
- The following phrases are extracted using spaCy.
  - VERB
  - PLACE
  - SUBJECT

Aggregation Phase

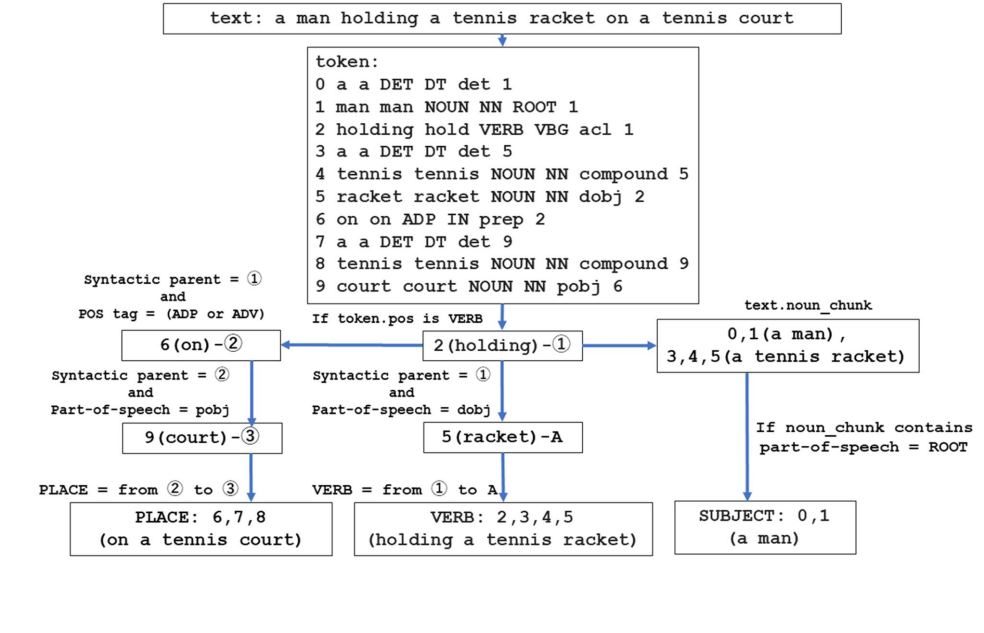


Figure 2. Flow of phrase extraction

Result

- it is thought that confirming the effectiveness of the method o f using OFA for captioning.
- the effectiveness of the word-by-word Lexrank method was not confirmed.

Table1. Names and methods of runs

Name	Caption generation	Caption aggregation
kslab_NUT_1	OFA	Sentence-Based
kslab_NUT_2	NIC	Sentence-Based
kslab_NUT_3	OFA	Word-By-Word
kslab_NUT_4	NIC	Word-By-Word

Table2. Scores for each run

	BLEU	CIDEr	CIDEr-D	METEOR	spice
kslab_NUT_1	0.1142	0.619	0.194	0.2764	0.097
kslab_NUT_2	0.0749	0.163	0.048	0.1971	0.049
kslab_NUT_3	0.0928	0.51	0.11	0.2217	0.071
kslab_NUT_4	0.0692	0.141	0.027	0.1642	0.036

Observations

- word-by-word Lexrank suggested that they failed to correctly recognize the sentence structure.
  - The object of transitive verbs sometimes cloud not be obtained.
- In the keyframe extraction phase, frames in which the subject was not clearly visible were sometimes selected.

Name	Caption generation	Caption aggregation
kslab_NUT_1	OFA	Sentence-Based
kslab_NUT_2	NIC	Sentence-Based
kslab_NUT_3	OFA	Word-By-Word
kslab_NUT_4	NIC	Word-By-Word

	BLEU	CIDEr	CIDEr-D	METEOR	spice
kslab_NUT_1	0.1142	0.619	0.194	0.2764	0.097
kslab_NUT_2	0.0749	0.163	0.048	0.1971	0.049
kslab_NUT_3	0.0928	0.51	0.11	0.2217	0.071
kslab_NUT_4	0.0692	0.141	0.027	0.1642	0.036

Conclusion

- Using OFA, which is more accurate in the caption phase, will increase the final accuracy.
- Word-by-word Lexrank was not much more accurate than sentence-based lexrank.