Waseda_Meisei at TRECVID 2018:Fully-automatic Ad-hoc Video Search

Yu Nakagome[1], Kazuya Ueki[1,2], Koji Hirakawa[1], Kotaro Kikuchi[1], Yoshihiko Hayashi[1], Tetsuji Ogawa[1], and Tetsunori Kobayashi[1]
[1] Waseda University, [2] Meisei University

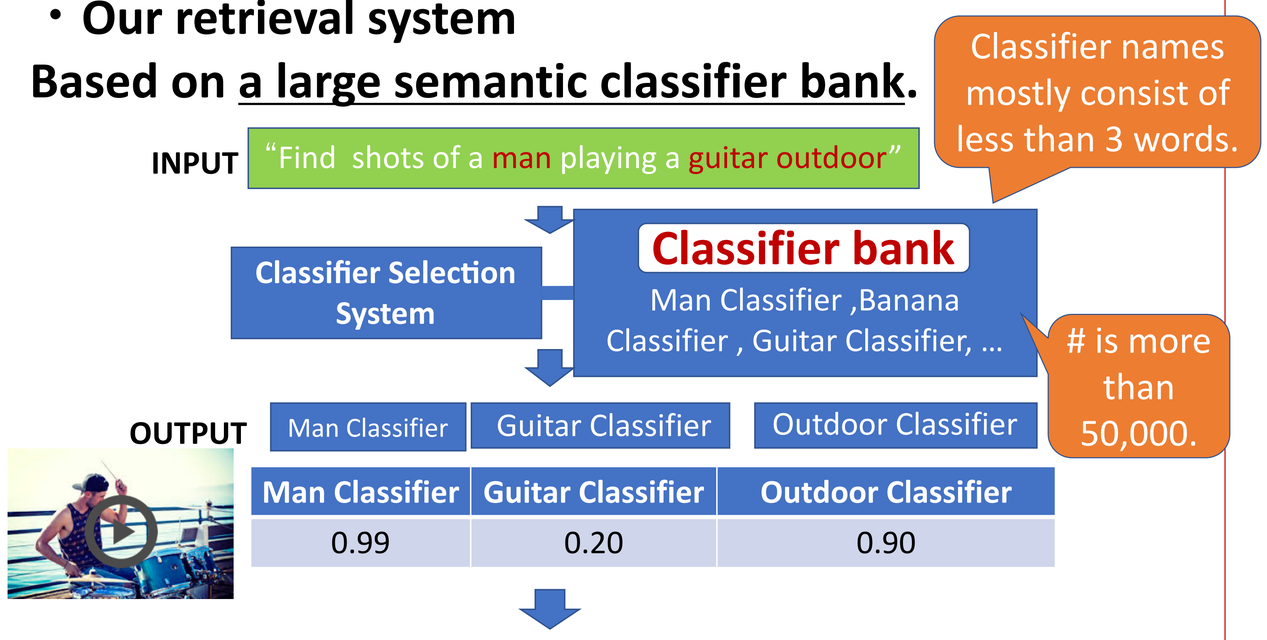
1.Background

- **Ad-hoc Video Search task objective:**

To return a list of at most 1000 shot IDs ranked according to their likelihood for each query.

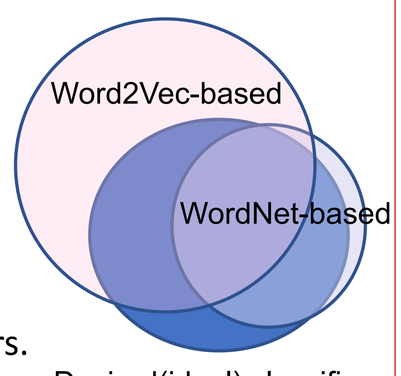
- **Our retrieval system**

Based on a large semantic classifier bank.



- **Conventional classifier selection methods**

- **Word2Vec-based method**
  - Based on cosine similarity of Word2Vec between keywords of query and classifier names.
- **WordNet-based method**
  - choosing classifiers if keywords of query and **synsets** of classifiers match.
- Classifier bank did not have much action classifiers.



Desired(ideal) classifiers

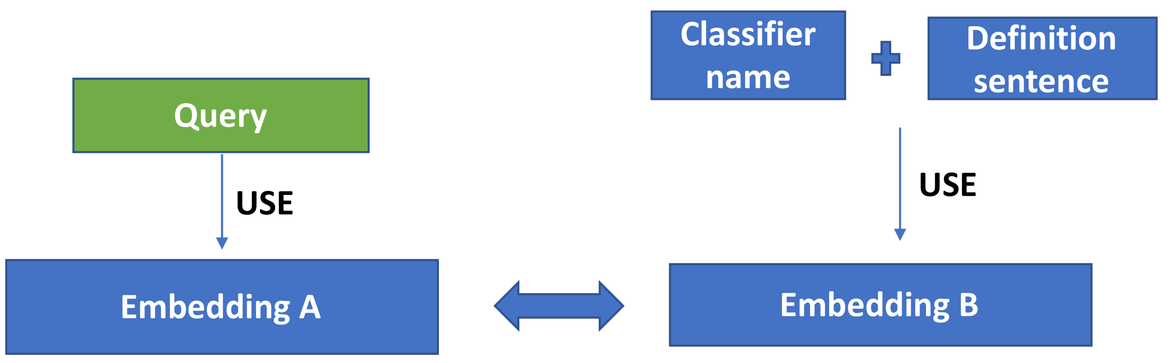
Problems to be solved.
❶ It is difficult to select appropriate classifiers by using just only classifier names or its synsets.
❷ We did not use action phrase in query.

2. Approach

①Using dictionary definition sentences for classifiers which trained by ImageNet

The purpose of this method is to select more appropriate classifiers by adding **auxiliary information** to classifier names.
We obtained **a vector representation of a sentence** by using **Universal Sentence Encoder (USE)** [Daniel Cer et al. 2018].

The **ImageNet** classes are linked to WordNet.
So the dictionary definition sentence is available.
e.g. S: (n) **smoker**, tobacco user (a person who smokes tobacco)

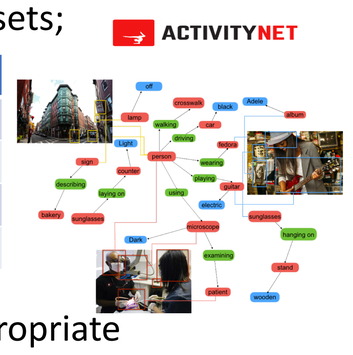


Cosine similarity; $\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\cdot\sqrt{\sum_{i=1}^{|V|} d_i^2}}$

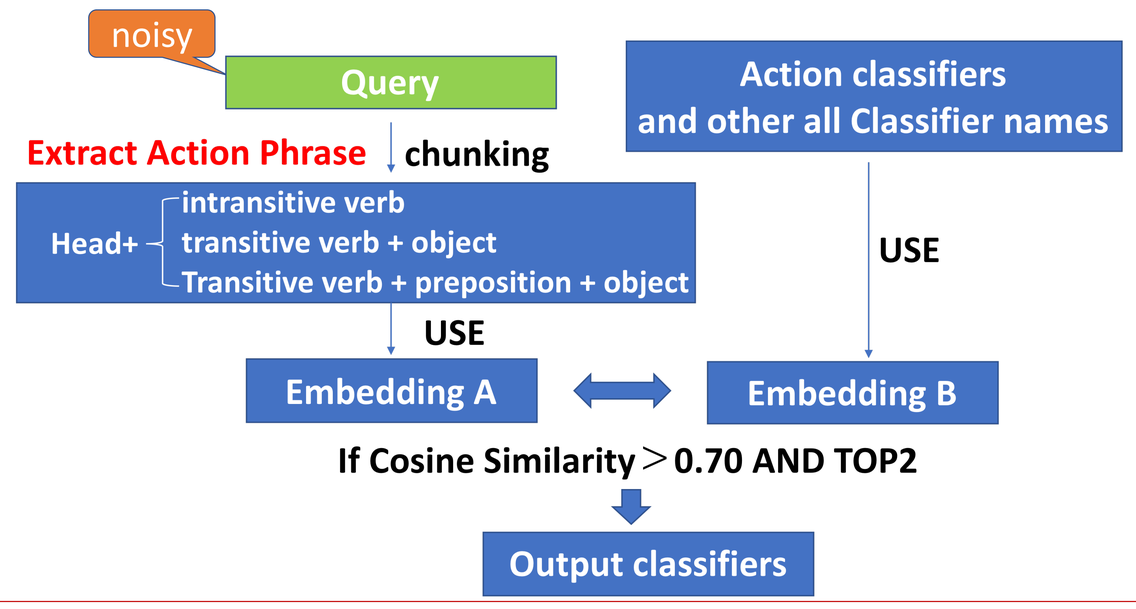
②Action recognition system for video search

We added action classifiers trained by following datasets;

Dataset	description
ActivityNet	Action, 200 class,648 video hours
Kinetics	Action, 400 class, 500,000 videos
Visual genome attribute	Adjective + noun
Visual genome relationship	Phonetic noun phrase (E.g. wearing a suit)



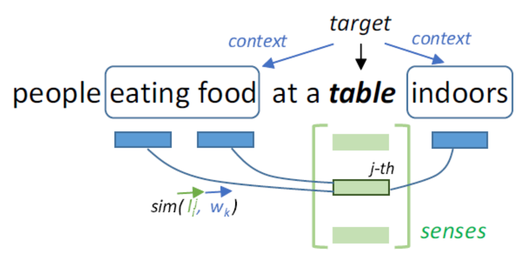
We proposed the following system which selects appropriate classifiers from the verb phrase in query .



3.Experiment

4.1 Experimental setup

- TRECVID 2018 AVS task.
- # of queries is 30.
- # of target videos is 335,994.
- Baseline is [Hirakawa et al.2017]
- Evaluation measure was set to MAP (Mean Average Precision)



4.2 Experimental results

①Using dictionary definition sentences for classifiers which trained by ImageNet

Method	MAP score
Baseline	0.0298
Baseline + Definition Weight(1:20)	**0.0424**

("Definition" is proposed system)

②Action recognition system

Method	MAP score
Manual (upper limit)	0.106
Baseline	0.0298
Action	0.0452
Baseline + Action (weight 1:50)	**0.0622**

("Action" is proposed system)

Query: Find shots of a person playing keyboard and singing indoors
->S: (n) **keyboardist** (a musician who plays a keyboard instrument)
Query: Find shots of a person lying on a bed
->S: (n) **rester** (a person who rests)
It was confirmed that both **proposed systems achieves good performance** in this task.

4. Results of Submitted Runs

method	MAP score
1.0*Definition + 99.0*Action+1.0* Baseline	**0.060**
1.0* Definition + 5.0*Action + 1.0*Baseline	0.057
1.0*Definition + 1.0* Action + 1.0*Baseline	0.040
2.0* Action + 1.0*Baseline	0.040
Manual(Upper limit)	**0.106**
13*definition + 30*Action + 1.0*Baseline (grid search)	**0.073**
RUCMM.18 (No.1 score)	**0.121**

Submitted runs

- Our team was **6th**.
- We combined three methods by calculating the weighted sum. The result changes greatly depending on this weight. Because the reliability of methods change with each query, it is difficult to define the weight.
- The problem is that even with the manual ideal classifier selection, we lose to other teams. We have to review classifier bank based system. Image recognition system should be updated by using object detection, scene graph.