Waseda_Meisei at TRECVID 2017

Ad-hoc Video Search(AVS)

Kazuya UEKI Koji HIRAKAWA Kotaro KIKUCHI
Tetsuji OGAWA Tetsunori KOBAYASHI

Waseda University
Meisei University

Highlights

- **AVS’s task objective：**

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

- **Our system:**

  Based on a large semantic concept bank.
  (More than 50,000 concepts)

- **This is our first submission to full automatic run:**

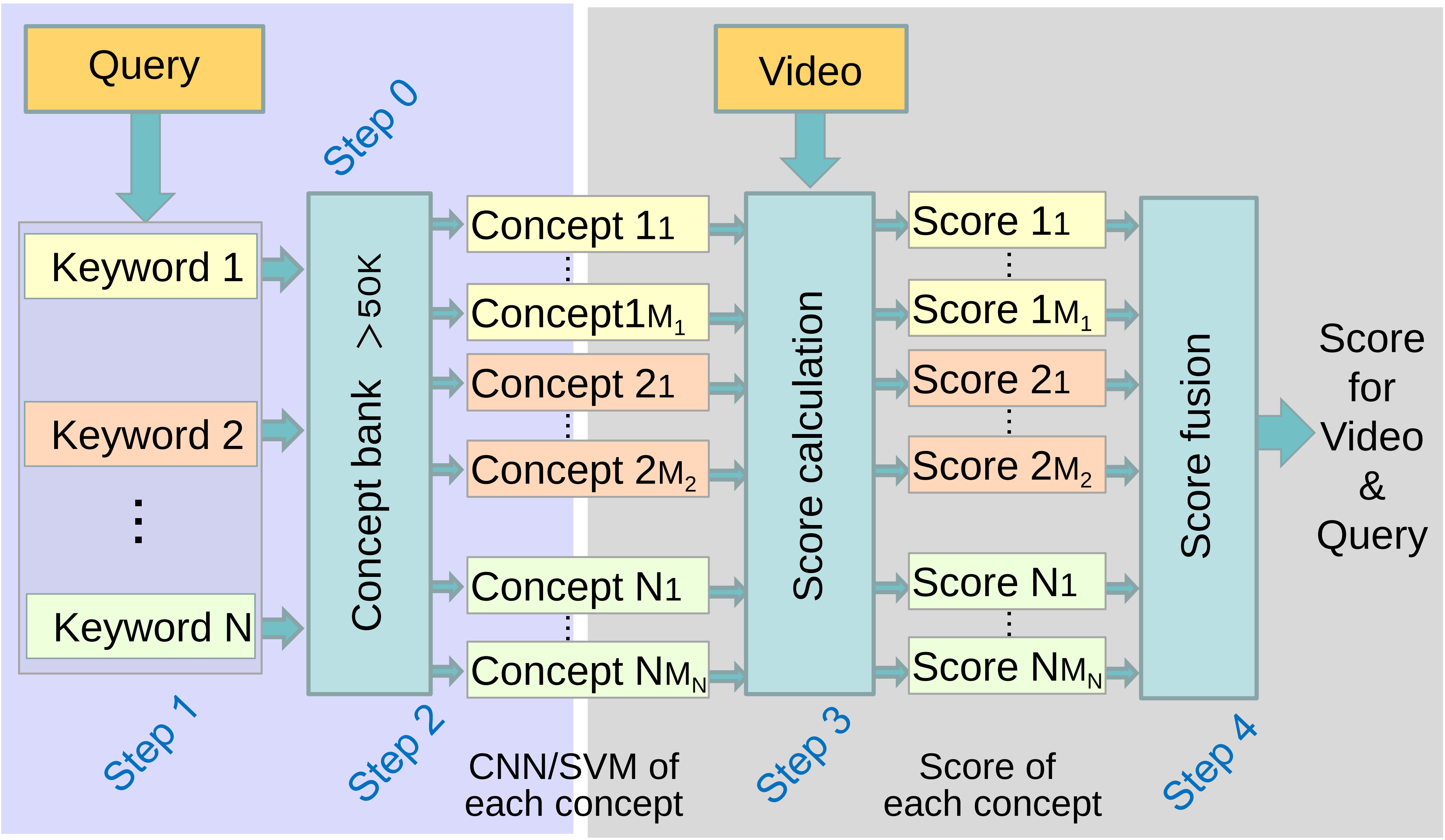
  Problem: Word ambiguity in concept selection step.
  WordNet/Word2Vec-based methods were proposed.
  WordNet-based one outperformed Word2Vec-based one.

1. System outline

1. System outline

New

Same as 2016 system



Query

Step 0

Keyword 1

Keyword 2

⋮

Keyword N

Step 1

Concept bank ＞50K

Step 2

Concept 1_1

⋮

Concept1_{M_1}

Concept 2_1

⋮

Concept 2_{M_2}

Concept N_1

⋮

Concept N_{M_N}

CNN/SVM of each concept

Video

Score calculation

Step 3

Score 1_1

⋮

Score 1_{M_1}

Score 2_1

⋮

Score 2_{M_2}

Score N_1

⋮

Score N_{M_N}

Score of each concept

Score fusion

Step 4

Score for Video & Query

Training Dataset

Training Dataset	Type	#Concepts, Data	Network	Model
TRECVID346 (ImageNet)	Object, Scene, Action	346 concepts	GoogLeNet	CNN/SVM tandem
PLACES205	Scene	205 concepts 2500K pictures	AlexNet	CNN
PLACES365	Scene	365 concepts 1800K pictures	GoogLeNet	CNN
Hybrid1183 (Places+ImageNet)	Object, Scene	1183 concepts 3600K pictures	AlexNet	CNN
ImageNet1000	Object	1000 concepts 1200K pictures	AlexNet	CNN
ImageNet4000,4437, 8201,12988	Object	4000,4437,8201, 12988 concepts	GoogLeNet	CNN
ImageNet21841	Object	21841 concepts 14200K pictures	GoogLeNet	CNN
FCVID239 (ImageNet)	Object, Scene,Action	239 concepts 91223 movies	GoogLeNet	CNN/SVM tandem
UCF101 (ImageNet)	Action	101 concepts 13320 movies	GoogLeNet	CNN/SVM tandem

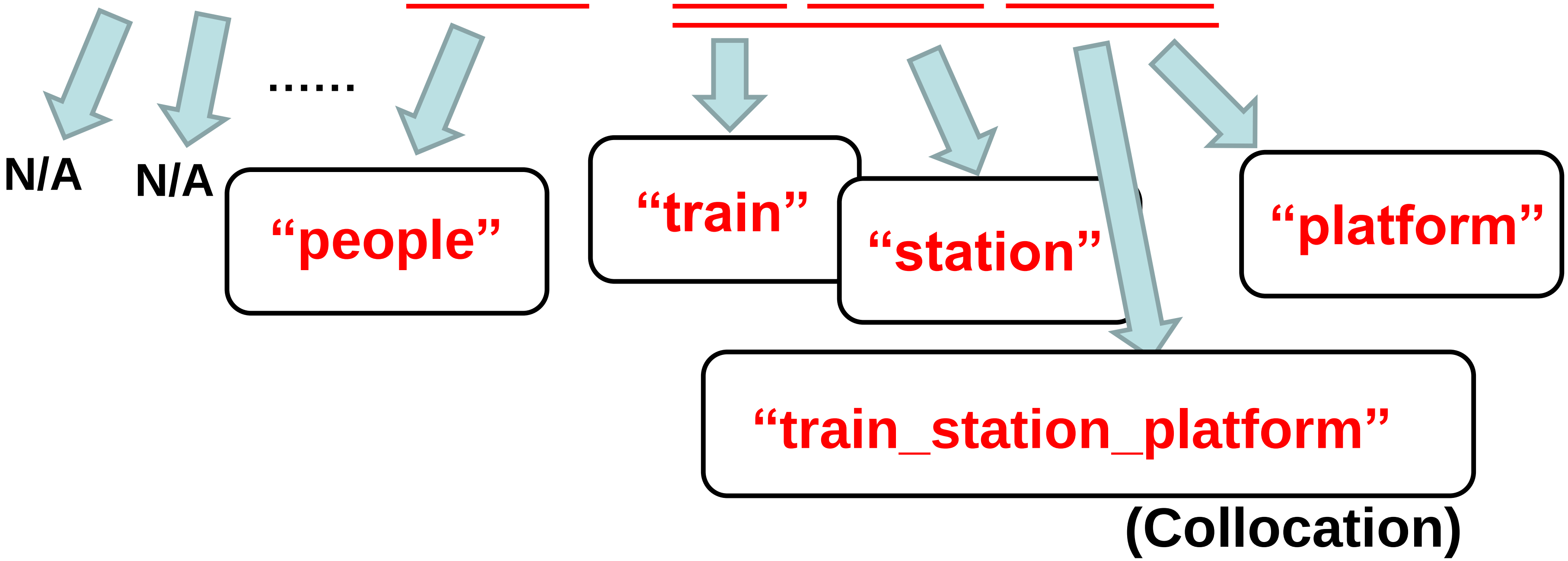
2. Detail of concept selection

2. Detail: Step 1 Extract keyword

Search keyword from query.

Query:

“One or more people at train station platform”



N/A

N/A

……

“people”

“train”

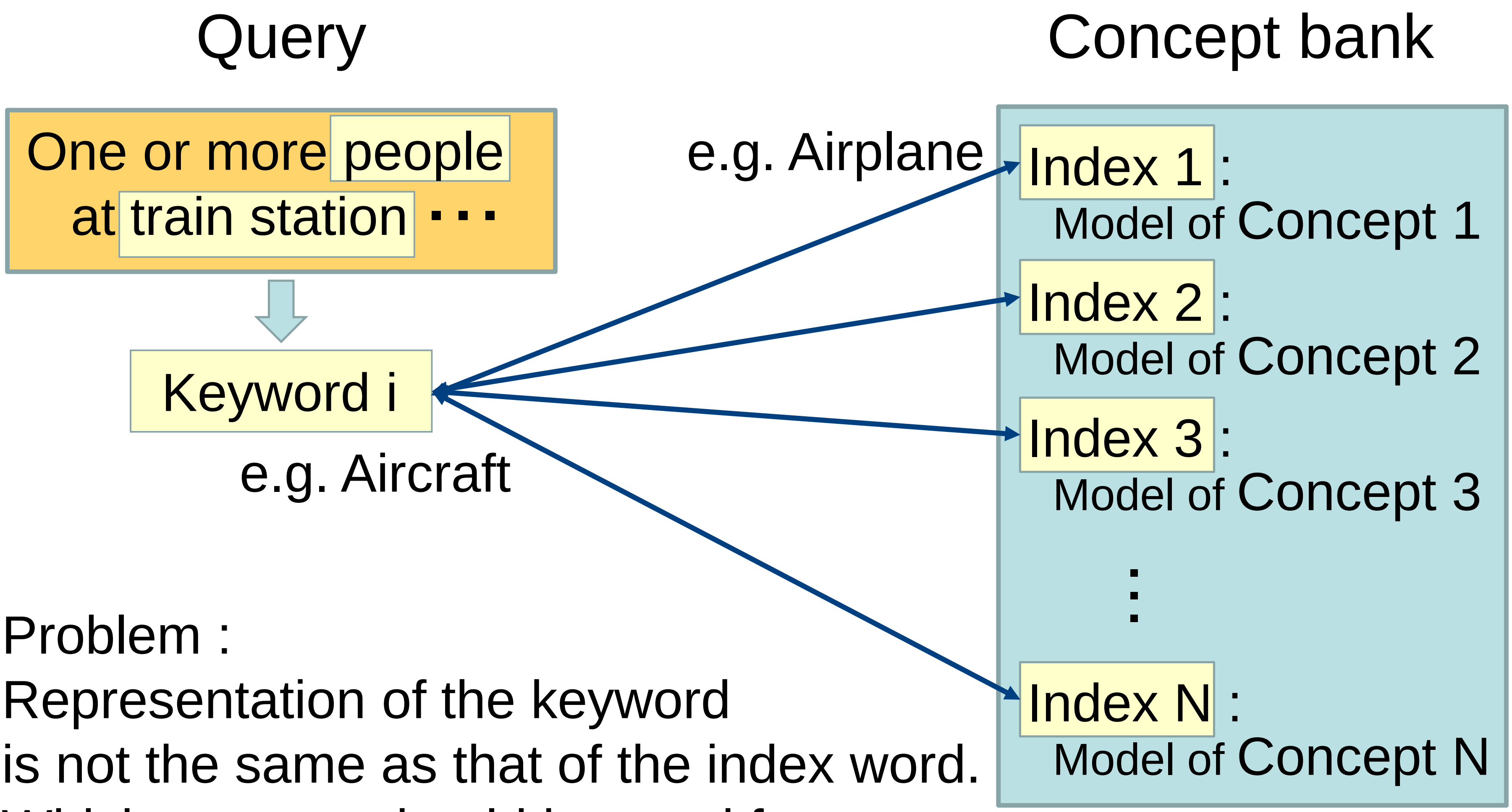
“station”

“platform”

“train_station_platform”

(Collocation)

2. Detail: Step 2 Choose concepts for each keyword



Query

One or more people at train station ･･･

Keyword i

e.g. Aircraft

e.g. Airplane

Concept bank

Index 1 : Model of Concept 1

Index 2 : Model of Concept 2

Index 3 : Model of Concept 3

⋮

Index N : Model of Concept N

Problem :
Representation of the keyword
is not the same as that of the index word.
Which concept should be used for
the keyword.

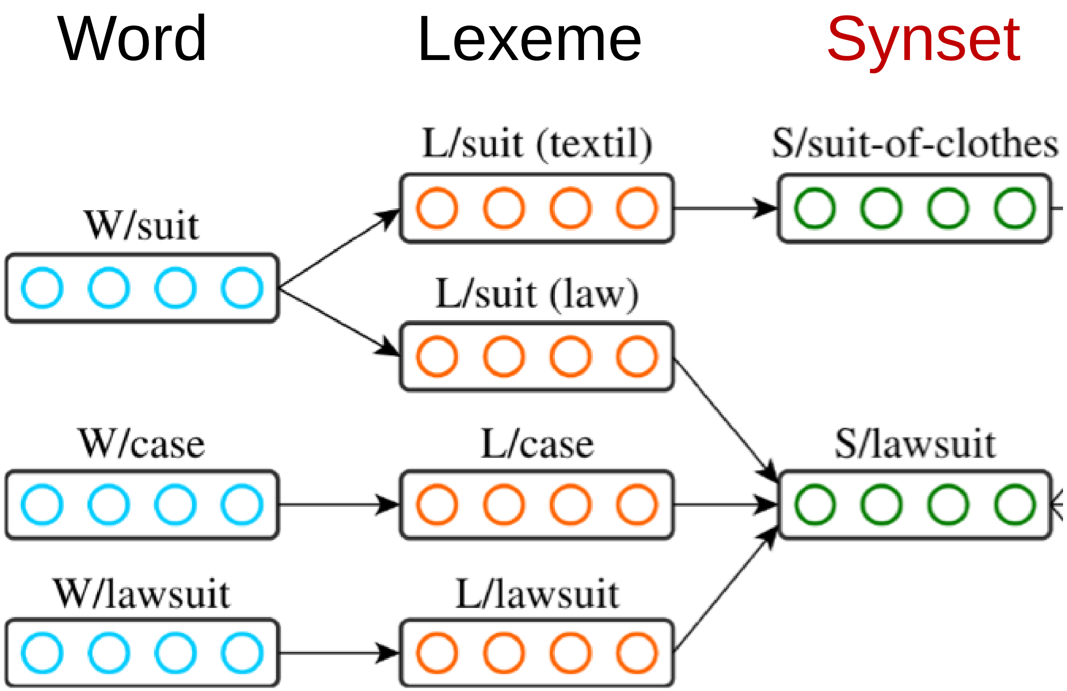
2. Detail: Step 2 Choose concepts for each keyword

- Manual runs
  - – The concept for the keyword is manually selected.

- Automatic runs
  - – WordNet based method.
    - Exact match of *synset*.
  - – Word2Vec based method.
    - Similarity of skipgram.
  - – Hybrid of WordNet & Word2Vec.

Automatic approach #1: WordNet *synset* matching

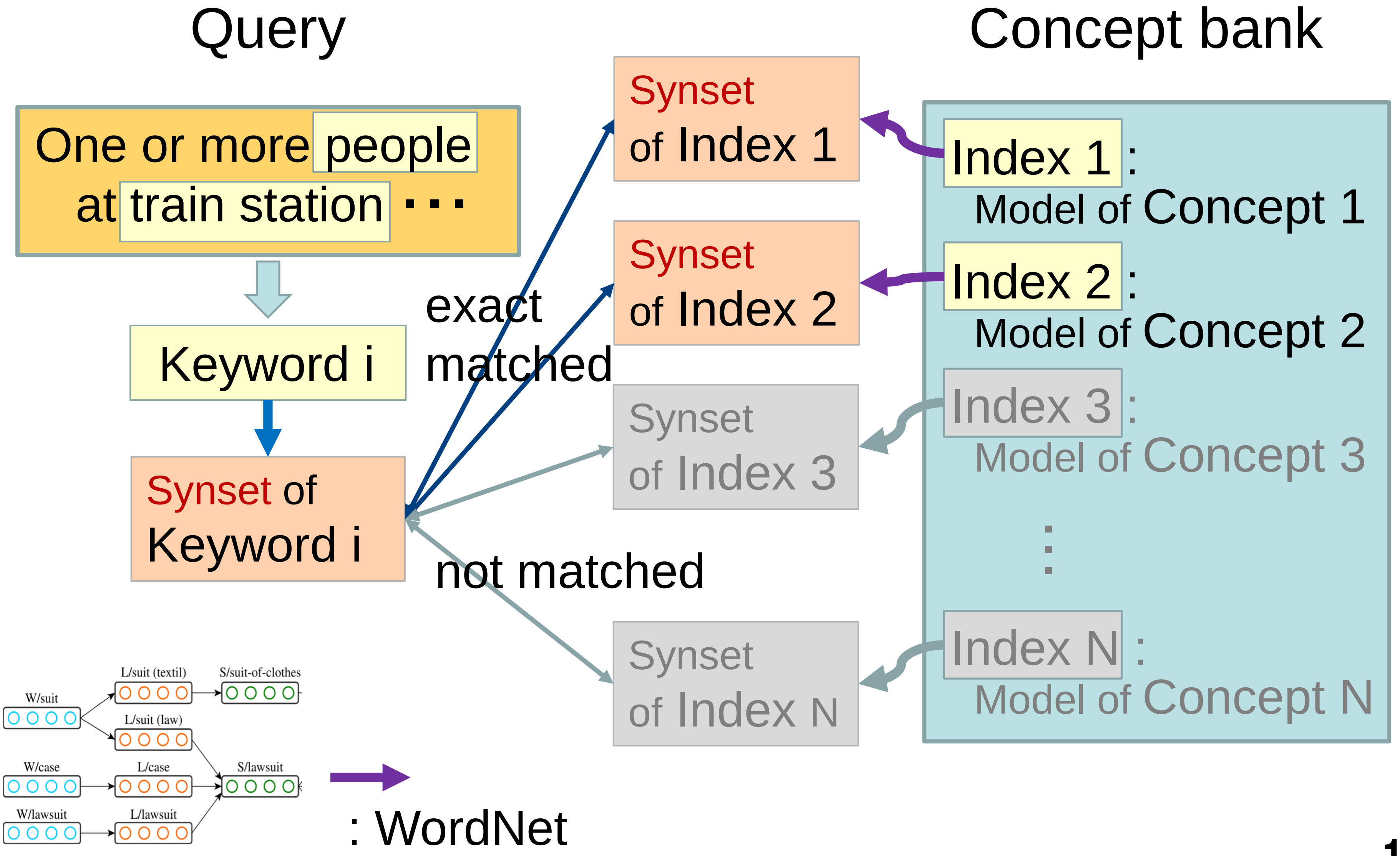
WordNet



Each “Word” has a set of “Lexeme”s.
Lexemes which have the same meaning make sysnset.

2. Detail: Step 2 Choose concepts for each keyword

Automatic approach #1: WordNet *synset* matching

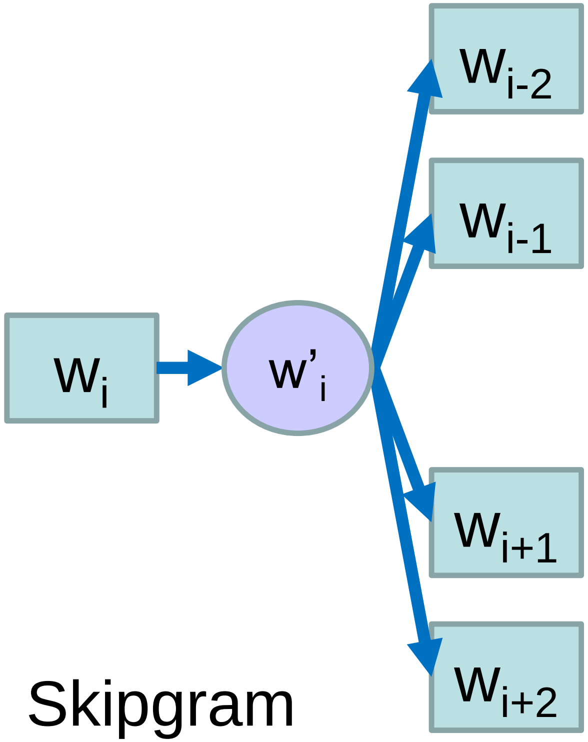


→ : WordNet

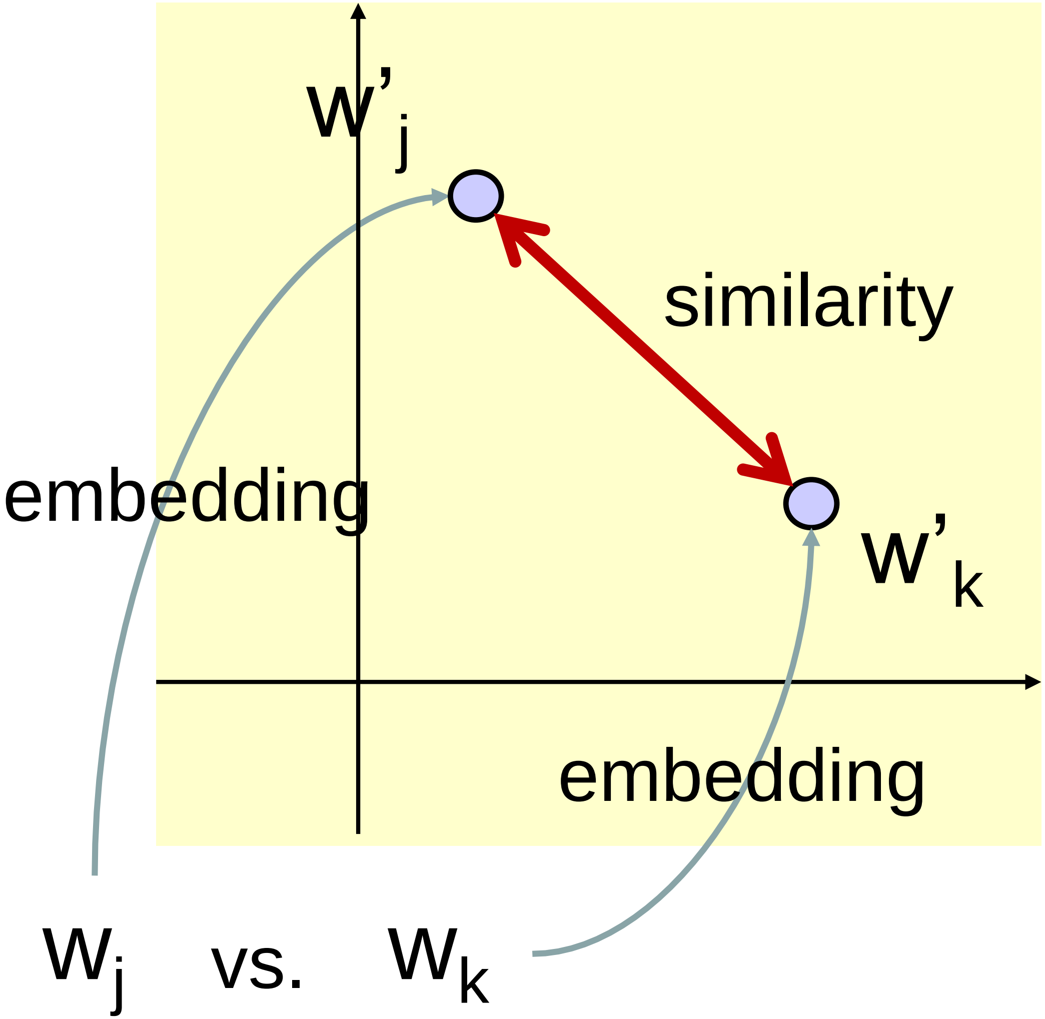
2. Detail: Step 2 Choose concepts for each keyword

Automatic approach #2: Word2Vec similarity

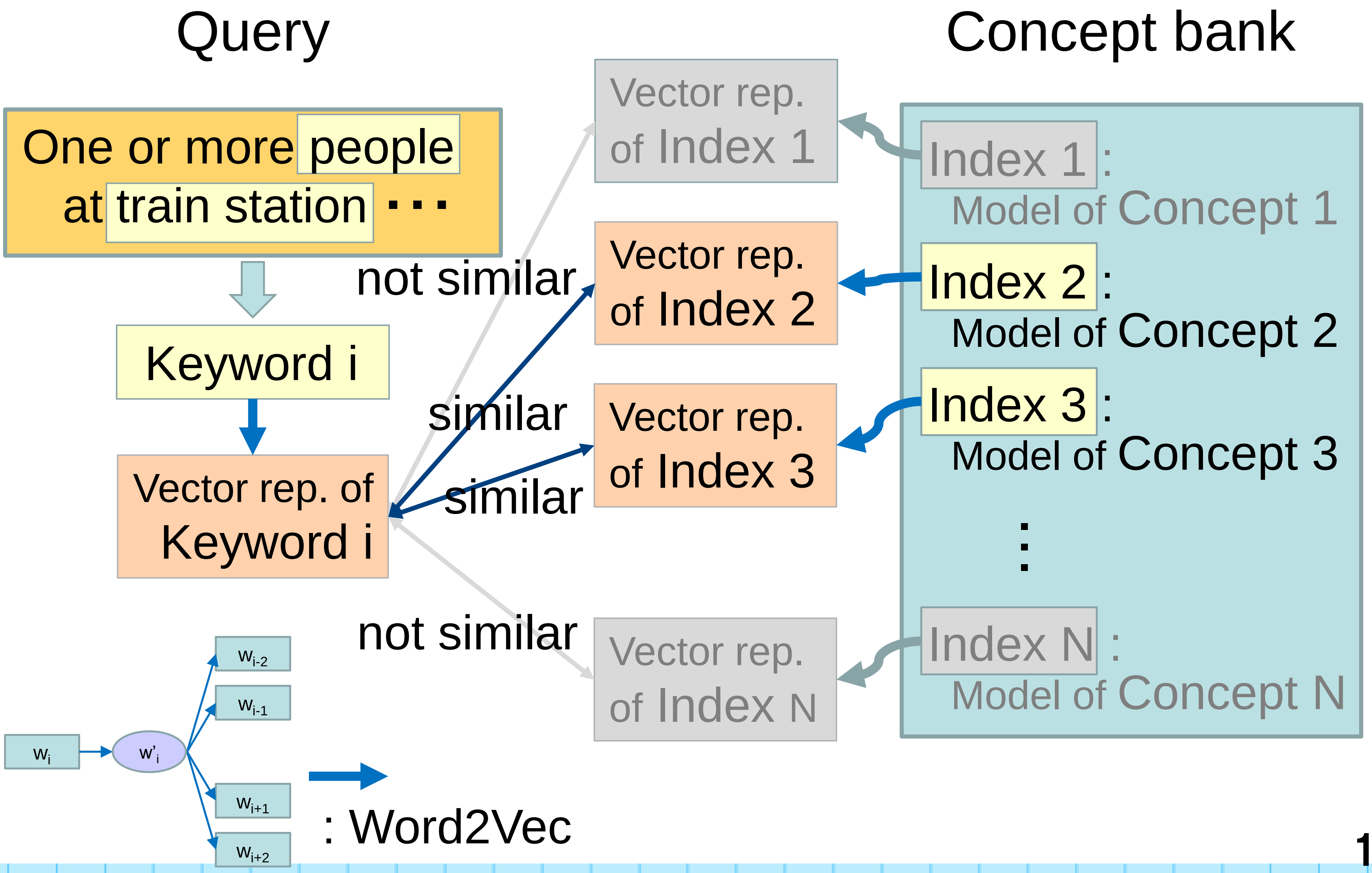
Word2Vec



Skipgram



Automatic approach #2: Word2Vec similarity



Query

Concept bank

One or more people at train station ･･･

Keyword i

Vector rep. of Keyword i

not similar

similar

similar

not similar

Vector rep. of Index 1

Vector rep. of Index 2

Vector rep. of Index 3

Vector rep. of Index N

Index 1 : Model of Concept 1

Index 2 : Model of Concept 2

Index 3 : Model of Concept 3

⋮

Index N : Model of Concept N

w_i

w'_i

w_{i-2}

w_{i-1}

w_{i+1}

w_{i+2}

: Word2Vec

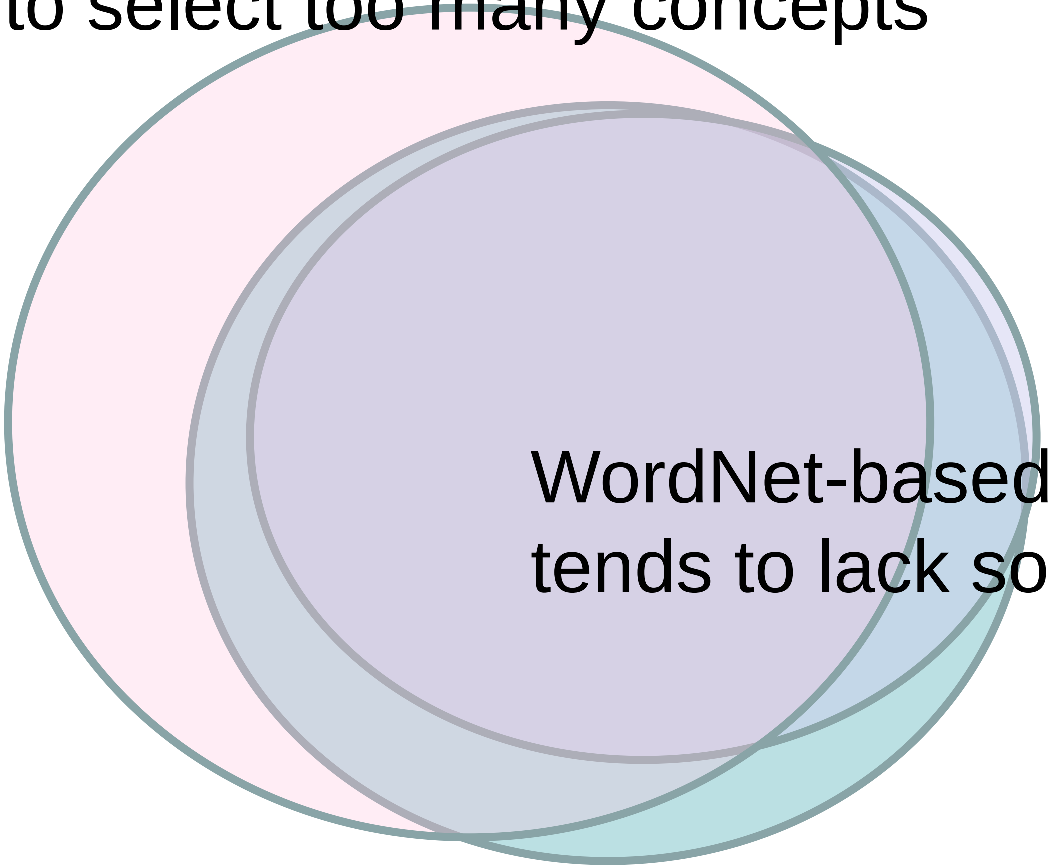
Automatic approach #3: Hybrid

Hybrid method:

```
Apply WordNet-based method, first.
If  failed /* WordNet-based method find no concepts */
    then Apply Word2Vec-based one.
```

2. Detail: Step 2 Choose concepts for each keyword

Expected Coverage



Word2Vec-based approach
tends to select too many concepts

WordNet-based approach
tends to lack some concepts.

Desired(ideal) Concept Set

2. Detail: Step 2 Calculate score

- TRECVID346
- FCVID239
- UCF101

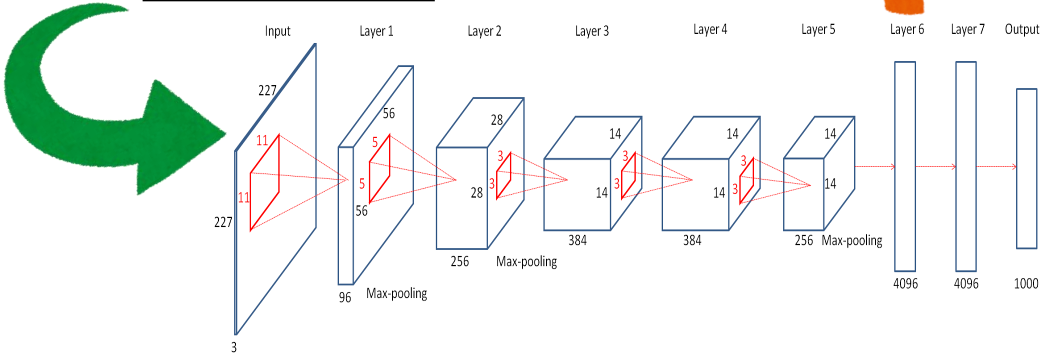
CNN/SVM tandem connectionist architecture

1st frame 2nd frame 10th frame

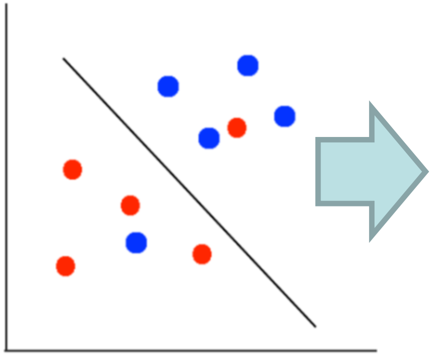
$$\begin{pmatrix} 2.051 \\ -1.349 \\ \vdots \\ \vdots \\ 2.493 \end{pmatrix} \begin{pmatrix} -9.251 \\ -3.039 \\ \vdots \\ \vdots \\ 1.455 \end{pmatrix} \cdots \begin{pmatrix} -3.482 \\ -1.498 \\ \vdots \\ \vdots \\ 2.411 \end{pmatrix} \xrightarrow{\text{max pooling}} \begin{pmatrix} 2.051 \\ -0.148 \\ \vdots \\ \vdots \\ 5.471 \end{pmatrix}$$

at most 10 images

hidden layer



CNN



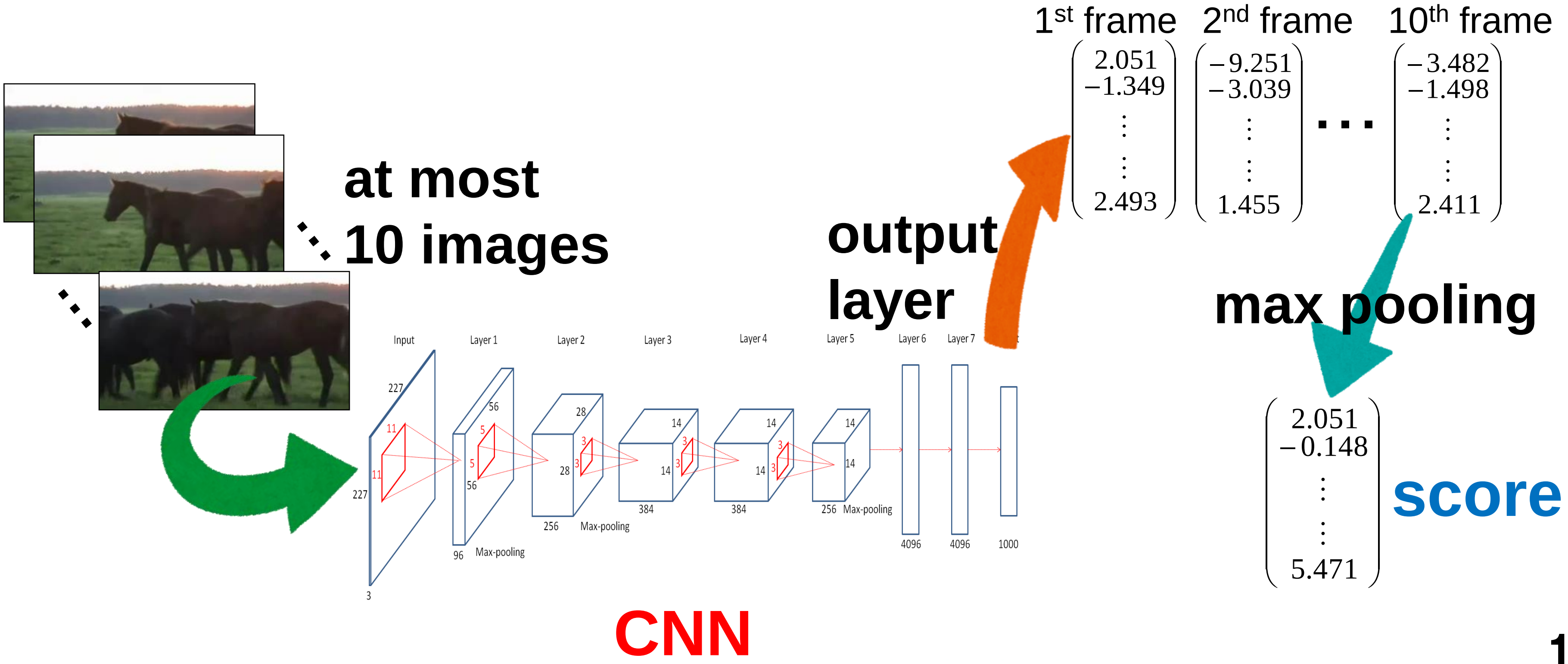
score

SVM

2. Detail: Step 2 Calculate score

PLACES205
PLACES365
HYBRID1183
IMAGENET1000
IMAGENET4000
IMAGENET4437
IMAGENET8201
IMAGENET12988
IMAGENET21841

The shot scores were obtained directly from the output layer (before softmax was applied)



$$\begin{pmatrix} 2.051 \\ -1.349 \\ \vdots \\ \vdots \\ 2.493 \end{pmatrix} \begin{pmatrix} -9.251 \\ -3.039 \\ \vdots \\ \vdots \\ 1.455 \end{pmatrix} \cdots \begin{pmatrix} -3.482 \\ -1.498 \\ \vdots \\ \vdots \\ 2.411 \end{pmatrix}$$

$$\begin{pmatrix} 2.051 \\ -0.148 \\ \vdots \\ \vdots \\ 5.471 \end{pmatrix}$$

3. Results

3. Results (Manual runs)

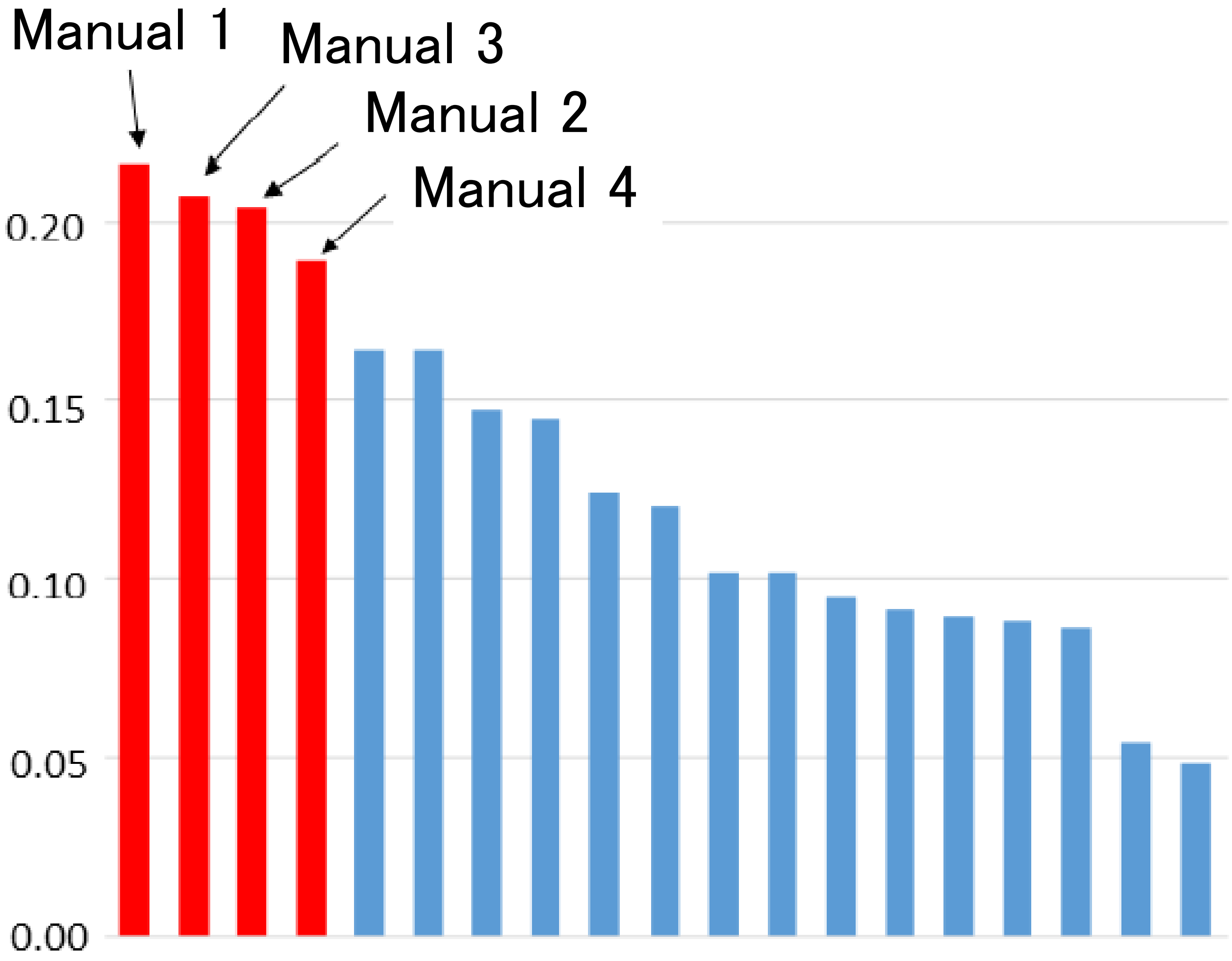
Comparison of Waseda_Meisei manual runs

Name	Fusion method	Fusion weight	mAP
Manual-1	Multiply(log)	✓	21.6
Manual-2	Multiply(log)		20.4
Manual-3	Sum(linear)	✓	20.7
Manual-4	Sum(linear)		18.9

Fusion method: **Multiply(log) > Sum(linear)**

Fusion weight: **w/ weight > w/o weight**

3. Results (Manual runs)



Comparison of Waseda Meisei runs with the runs of other teams for all submitted manually assisted runs.

3. Results (Automatic runs)

Comparison of Waseda_Meisei automatic runs

Name	WordNet synset	Word2Vec	FCVID239 ＋UCF101	mAP
Auto-1	✓			15.9
Auto-2		✓		14.3
Auto-3		✓	✓	14.1
Auto-4	✓	✓		12.5

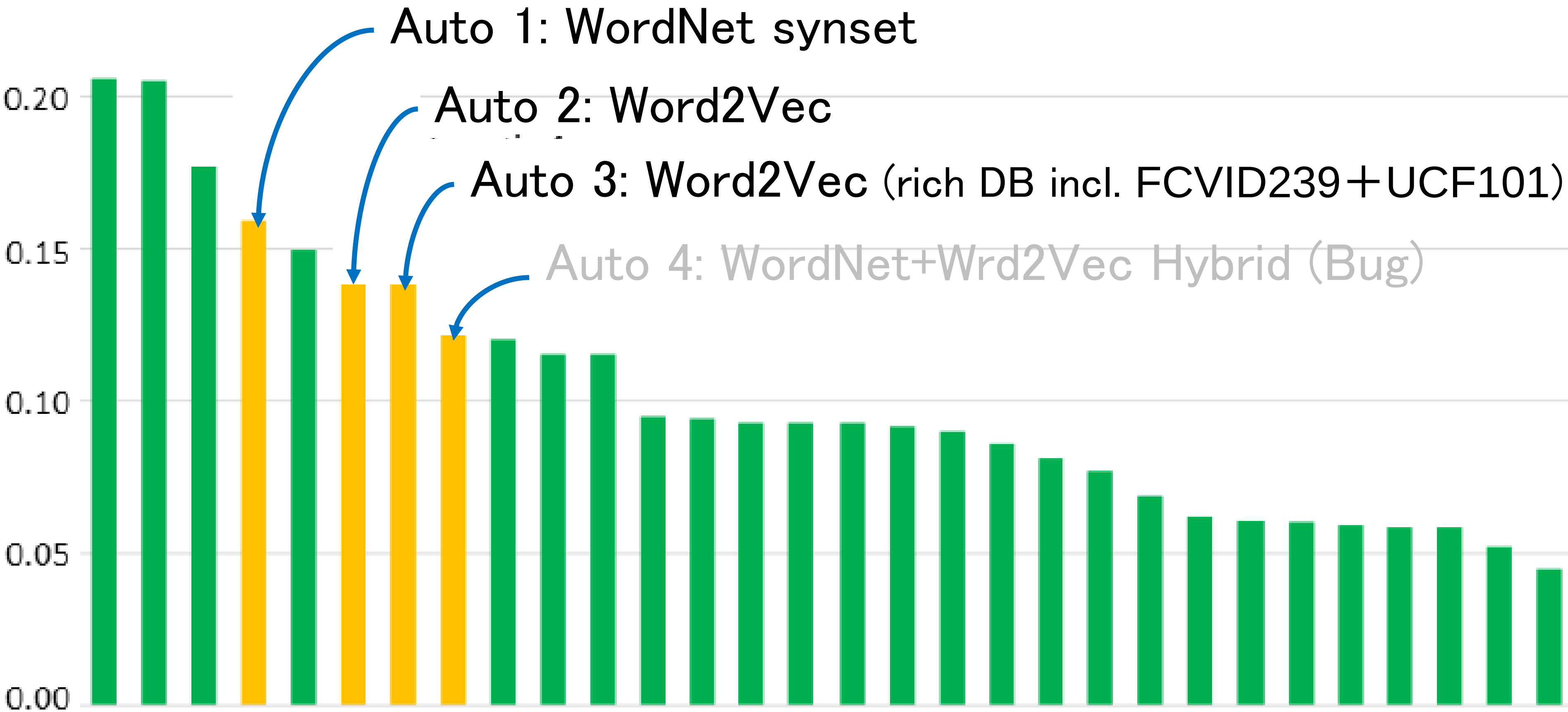
WordNet vs. Word2Vec: WordNet > Word2Vec

3. Results

Results for 2016 TRECVID dataset

Name	WordNet synset	Word2Vec	FCVID239 +UCF101	mAP
Auto-1	✓			17.8
Auto-2		✓		17.4
Auto-3		✓	✓	17.4
Auto-4	✓	✓		17.8

3. Results (Automatic runs)



Auto 1: WordNet synset

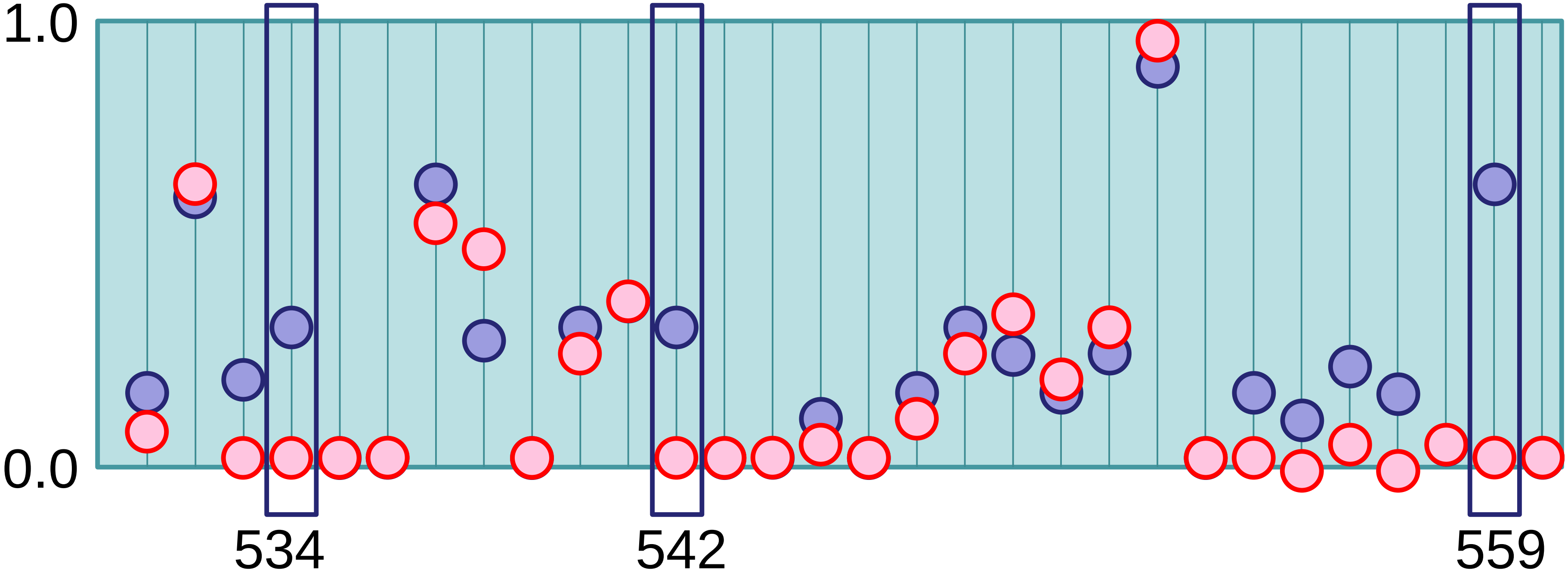
Auto 2: Word2Vec

Auto 3: Word2Vec (rich DB incl. FCVID239＋UCF101)

Auto 4: WordNet+Wrd2Vec Hybrid (Bug)

Comparison of Waseda Meisei runs with the runs of other teams for all the fully automatic runs.

3. Results: Difference btw. our Auto○ & our Manu.●

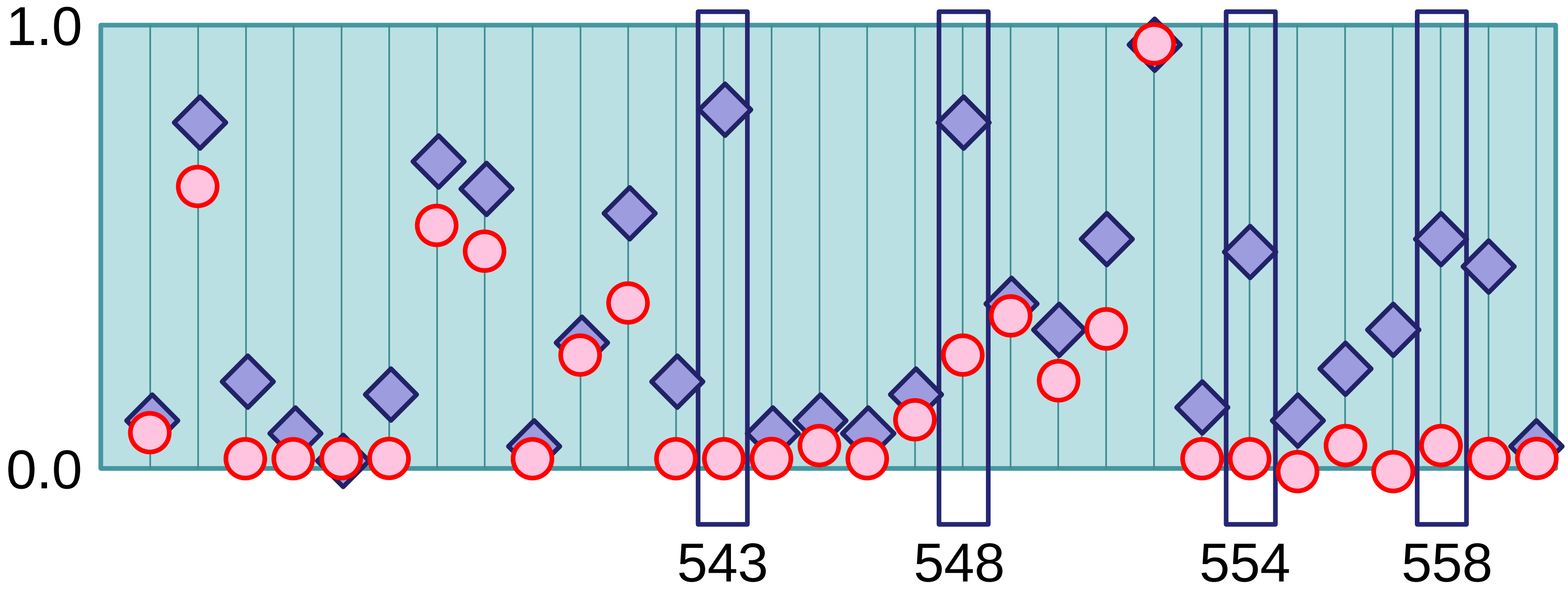


534 Find shots of ***a person talking behind a podium*** wearing a suit outdoors during daytime → "Speaker_At_Podium" is used in manu.

542 Find shots of ***at least two*** planes both visible
→ Object counting module is installed in manual condition.

559 Find shots of a man and woman ***inside a car***
→ "car_interior" is used and "car" is not used in manual.
(All, parsing (linguistic) problem)

3. Results: Difference btw. our Auto○ & Top◇



543 Find shots of a person communicating using ***sign language***

→ No concept for "sign language". (Short of concepts)

554 Find shots of a person holding or operating ***a TV or movie camera***

→ "TV" contaminated. (Parsing problem)

558 Find shots of a person wearing a ***scarf***

→ "scarf_joint" contaminated. (Word-concept matching problem)

Scarf itself is difficult to recognize. (Scoring problem)

4. Summary & future works

4. Summary and future works

Summary

- We joined in “ad-hoc video search” task.

- This is our first attempt to “automatic run”.
  In step2 (selection of concepts from keyword),
  WordNet-based/Word2Vec-based methods proposed

- WordNet-based concept selection outperformed
  Word2Vec-based one.

4. Summary and future works

Future works

- To improve the concept selection methods.

  e.g. Other use of WordNet / Word2Vec

- To improve linguistic part.

  e.g. a person talking behind xxxx,
  inside car,
  at least two xxxx
  TV or movie camera

- To handle action type concepts.

Thank you for your attention.

Any questions?